Christian Reske (Heidelberg)

WebCorp Linguist's Search Engine (LSE) – eine übersetzungsbezogene Evaluation

Christian Reske (Heidelberg)

WebCorp Linguist's Search Engine (LSE) – eine übersetzungsbezogene Evaluation

Abstract:

The following article is focused on the evaluation of the *Linguist's Search Engine (LSE)*, a web-based corpus created by the Birmingham City University (BCU). Following an introduction of underlying principles and user interface of BCU's *LSE*, the latter is put to the test and compared with the *Corpus of Contemporary American English* (*COCA*), considering the applicability of both corpora in technical translation. Based on a test run using selected terms in English, suggestions for web-based multi-language and translation corpora are provided.

Der vorliegende Artikel untersucht die von der Birmingham City University entwickelte *Linguist's Search Engine* (*LSE*). Bei *LSE* handelt es sich um ein Korpus, dessen Daten dem Web entnommen sind. Nach einer kurzen Einführung in Funktionsweise und Benutzeroberfläche, wird *LSE* einem kleinen Testdurchgang unterzogen. Im Anschluss werden die Ergebnisse mit denen aus dem *Corpus of Contemporary English* (*COCA*) verglichen. Beide Korpora werden somit auf ihre Einsatzmöglichkeiten bei der Übersetzung technischer Texte evaluiert. Abschließend werden, unter Berücksichtigung der Testergebnisse, Vorschläge zu webbasierten, mehrsprachigen Korpora sowie zu Übersetzungskorpora thematisiert.

Keywords:

computational linguistics; corpus linguistics; translation of technical texts; usability; evaluation; the web as corpus

Computerlinguistik; Korpuslinguistik; technische Übersetzung; Usability; Evaluation; das Web als Korpus

Inhalt:

1 Verfügbare Korpora und deren Nutzen für die Praxis des Übersetzens

Mit der fortschreitenden Technisierung des Übersetzungs- und Rechercheprozesses spielen Korpora als übersetzungsbezogene Hilfsmittel eine bedeutende Rolle. Waren Korpora ursprünglich von Linguisten und später von Computerlinguisten als empirische Werkzeuge zur Untersuchung von Sprache(n) entwickelt worden, so sind sie heute auch für professionelle Übersetzer im täglichen Gebrauch unersetzlich geworden. Als Hilfsmittel zur Überprüfung von Kollokationen tragen sie zur sprachlichen Qualität des übersetzten Textes bei, vor allem dann, wenn es sich um die Verifikation von Formen des gebundenen Sprachgebrauchs in den Fremdsprachen handelt. Neben (elektronischen) Wörterbüchern und Glossaren sollten Korpora demnach zur Grundausstattung eines Übersetzers gehören. Internet-Plattformen zur Korpuserstellung und -verwendung sind im deutschen und englischen Sprachraum weit verbreitet. Hier stehen Übersetzern und Forschern gleich mehrere Korpora zur Auswahl. Bedingt durch die verschiedenen Korpora variieren auch die Definitionen. Für diesen Artikel soll folgende Definition gelten:

> Ein Korpus ist eine Sammlung schriftlicher oder gesprochener Äußerungen in einer oder mehreren Sprachen. Die Daten des Korpus sind digitalisiert, d.h. auf Rechnern gespeichert und maschinenlesbar. Die Bestandteile des Korpus, die Texte oder Äußerungen, bestehen aus den Daten selbst sowie möglicherweise aus Metadaten, die diese Daten beschreiben, und aus linguistischen Annotationen, die diesen Daten zugeordnet sind. Wenn wir von linguistischen Korpora sprechen, dann handelt es sich um Textsammlungen mit kompletten Texten oder zumindest mit sehr großen Textausschnitten. Außerdem sollten linguistische Korpora meist repräsentativ, durch Metadaten erschlossen und linguistisch annotiert sein. (Lemnitzer und Zinsmeister 2006: 40-41)

Für den deutschen Sprachraum stehen, neben einigen Spezialkorpora, das *Digitale Wörterbuch der deutschen Sprache* der Berlin-Brandenburgischen Akademie der Wissenschaften (*DWDS*)[1], *COSMAS II* und der *Wortschatz der Universität Leipzig* zur Verfügung. Zu den englischsprachigen Korpora zählen u. a. das *British National Corpus* (*BNC*) und das *Corpus of Contemporary American English* (*COCA*). Diesen Korpora ist gemein, dass

[1] Das *DWDS*-Projekt beinhaltet ein eigenes Referenzkorpus der deutschen Sprache des 20. und frühen 21. Jahrhunderts, das für Recherchen zur Verfügung steht. Darüber hinaus können über die Website des *DWDS* verschiedene externe Referenzkorpora, Zeitungskorpora und Spezialkorpora konsultiert werden.

sie über einen festen und im Vorfeld annotierten Datensatz verfügen. Dieser Datensatz muss erweitert, überprüft und gepflegt werden, um die Veränderungen in der Sprache und im Sprachgebrauch wiedergeben zu können. Bei den in die Korpora aufgenommenen Materialien handelt es sich vorwiegend um Texte aus Zeitungen, Zeitschriften und Büchern.

Welche Möglichkeiten stehen aber Fachübersetzern zur Verfügung, wenn mittels Korpora fachsprachliche Kollokationen recherchiert oder überprüft werden sollen? Im Zuge der Forschung im Bereich der maschinellen Sprachverarbeitung wurden Programme entwickelt, mit deren Hilfe Fachtermini und auch Kollokationen aus dem Internet extrahiert werden können. Auf diese Weise lassen sich für spezielle Übersetzungs- und Dolmetschaufträge Spezialkorpora aus dem Internet erstellen. Exemplarisch sei hier auf *BootCat* als Programm und auf die Arbeiten von Heid (2011) für die Anwendung und Forschung verwiesen.

Nicht jeder Benutzer möchte ein eigenes Programm für eine besondere Aufgabe auf seinem Rechner installieren. Steht man desktopbasierten Lösungen eher kritisch gegenüber, so eignen sich vor allem webbasierte Anwendungen. Die hier vorgestellte *Web Corp Linguist's Search Engine* (*WebCorp LSE*) wurde von der Birmingham City University als Webanwendung entwickelt.[2] Für die Benutzung ist keine Installation einer weiteren Software nötig. Wie auch bei *COCA* und den deutschsprachigen Korpora, handelt es handelt sich bei *WebCorp LSE* ebenfalls um ein linguistisches Forschungskorpus.

Im vorliegenden Beitrag soll die Tauglichkeit dieser *Linguist's Search Engine* (*LSE*) für eine übersetzungsbezogene Recherche untersucht werden. Ferner soll die *LSE* hinsichtlich Benutzerfreundlichkeit und Ausgabe bzw. Anzeige der Treffer mit dem *COCA* verglichen werden. Bei der Gegenüberstellung beider Korpora wurden keine allgemeinsprachlichen, sondern fachsprachliche Termini verwendet. Die Beispiele wurden einem vom Vf. erstellten Übersetzungskorpus zum Fachgebiet der Zementherstellung für das Sprachenpaar Deutsch / Englisch entnommen (Reske 2012). Da *COCA* nur für die englische Sprache verfügbar ist,

[2] Die hier vorgestellte Linguist Search Engine steht nicht in Verbindung mit dem von Aaron Elkiss entwickelten Programm, das ebenfalls unter dem Namen LSE in Suchmaschinen gelistet ist. Dieses Projekt ist seit 2010 eingestellt, Dokumentation und Berichte können unter http://lse.umiacs.umd.edu/ eingesehen werden.

wurde bei der Untersuchung beider Korpora auf die englischen Termini zurückgegriffen, *WebCorp LSE* lässt jedoch auch die Eingabe von deutschen Suchbegriffen zu. Die deutschen Ausgangstermini wurden zum besseren Verständnis in Klammern hinzugefügt. Auf eine detaillierte Beschreibung der Funktionsweise bzw. der Benutzeroberfläche des *COCA* wird verzichtet, es sei aber an dieser Stelle auf die von Mark Davis erstellte Dokumentation unter der Rubrik „Where should I start?“ und die übersetzungsbezogene Betrachtung von Bilić et al. (2009: 134-36) verwiesen.

Basierend auf den Erfahrungen aus Studium und Beruf vertritt der Vf. die These, dass sich die *WebCorp LSE* besser zur Kollokationsrecherche für Fachtexte eignet, während *COCA* bessere Ergebnisse für die Verifikation von Kollokationen in allgemeinsprachlichen Vertextungsprozessen liefert. Eine weitere Hypothese besteht darin, dass mit einem zunehmenden Fachlichkeitsgrad des Textes die Anzahl der Kollokationen abnimmt (vgl. Bilić et al. 2009: 130-32; Halkiopoulou 2006). Für den im Rahmen dieses Beitrags behandelten Vergleich wurden die unten stehenden Nomina aus einem vom Vf. erstellten Übersetzungskorpus herausgelöst und zu Testzwecken verfügbar gemacht. Die einzelnen Termini wurden nicht, wie bei linguistischen Forschungen üblich, nach ihrer Frequenz im vorhandenen Übersetzungskorpus entnommen. Die Auswahl der Termini erfolgte aufgrund besonderer Schwierigkeiten, die sich bei der Kollokationsrecherche und des damit verbundenen Übersetzungsprozesses ergaben (Reske 2012).

Diese Probleme lagen zum einen darin, dass es sich um Fachtermini handelt, die speziell dem Fachgebiet der Zementarten und der Zementherstellung zuzuordnen sind. Zum anderen handelt es sich um Termini, deren allgemeinsprachliche Bedeutung stark von der fachspezifischen abweicht (zur fachsprachlichen Lexik und linguistischer Grundlagen vgl. Fluck: 1996: 35-37 und 47-50).

Ein weiteres Problem stellen die Mehrworttermini dar. In Bezug auf die Mehrworttermini geht der Vf. davon aus, dass diese nur unzureichend bis gar nicht von den Suchsyntaxen der untersuchten Korpora als solche erkannt werden. Dies führt zu folgender, in dieser Untersuchung verwendeten Wortliste: *initial setting time* (Anfangserhärtung); *flexural strength* (Biegezugfestigkeit); *rotating kiln* (Drehrohrofen); *compressive strength*

(Druckfestigkeit); *aggregate* (Gesteinskörnung); *limestone marl* (Kalksteinmergel); *calcining zone* (Kalzinierzone); *wet process* (Nassverfahren); *raw meal* (Rohmehl); *pore formation* (Porenbildung), *transport cement* (Transportzement).

Für den Übersetzer sind Gedanken zur Tokenisierung und Quantifizierung in der beruflichen Praxis von geringerer Bedeutung. Alle Hilfsmittel müssen einfach zu bedienen und schnell verfügbar sein. Linguistische Bedenken über mögliche Einflüsse bestimmter Daten auf die Empirie zählen in der übersetzerischen Praxis erfahrungsgemäß wenig. Bei Kollokationen möchten die Übersetzer, den Ko- und den Kontext des Fachwortes überprüfen.

So stellt sich, mit Bezug auf die eingangs vorgestellten Termini, die Frage, welche Nachbarn beispielsweise dem Terminus *rotary kiln* im Sprachgebrauch zugeordnet sind. Bei allgemeinsprachlichen Texten wäre der erste Anlaufpunkt das *Oxford Collocations Dictionary* (McIntosh 2009), jedoch sind die Einträge hier aufgrund des Speichermediums (Buch bzw. CD-ROM) beschränkt. Fachsprachliche Kollokationen sind in diesem Werk nur in sehr geringem Umfang verzeichnet. Erfahrene Übersetzer greifen deshalb auf *COCA* zurück. Alle eingangs erwähnten Begriffe wurden daher vom Vf. jeweils in *COCA* und in *WebCorp LSE* überprüft.

Bevor *WebCorp LSE* mit *COCA* verglichen wird, sei kurz auf die Funktionsweise von *WebCorp LSE* eingegangen.

2 Prinzip und Aufbau von *WebCorp LSE*

Auch wenn es sich bei der *LSE* um ein webbasiertes Korpus handelt, bedeutet dies nicht, dass die Daten „irgendwo im Netz liegen“. Es handelt sich also demnach auch nicht um eine von und für Linguisten geschaffene Form des *Cloud-Computing*. Wie *COCA* verfügt auch *WebCorp LSE* über physisch abgespeicherte und annotierte Daten.

Im Unterschied zu *COCA* sind diese Daten jedoch ausschließlich dem Internet entnommen. Mit der Vorgängerversion der *WebCorp LSE*, dem sogenannten *WebCorp*, wurde erstmals der Versuch unternommen, aus dem WWW ein Korpus zu erstellen. Die Benutzeroberfläche des *WebCorp* unterscheidet sich nur geringfügig von denen der bekannten Suchmaschinen.

Der Benutzer kann sowohl einzelne Wörter als auch Sätze in das Suchfeld eingeben. Durch weitere Optionen kann die Suche eingegrenzt und/oder verändert werden. Technisch betrachtet ist *WebCorp* den Suchmaschinen vorgeschaltet, verwendet werden *BING*™ und *Google*™ . Die durch die Suchanfrage erstellte Liste, also die Treffer der Suchmaschinen, werden über spezielle Algorithmen von *WebCorp* als Konkordanzen ausgegeben. Die gesammelten Konkordanzen werden anschließend als *keywords in context* (*KWIC*) in einer einzelnen Liste übersichtlich dargestellt. Es werden ebenfalls die URLs der Webseite angezeigt. Über die Anzeige der URL können Übersetzer entscheiden, ob die gelistete Konkordanz aus einer verlässlichen Quelle stammt und somit für die Übersetzung angemessenen ist (vgl. Bilić 2009: 204-210). Für weitere Informationen zur Funktionsweise des *WebCorp* sei hier auf die Homepage dieses Projektes und auf Renouf (2009) verwiesen.

Soll jedoch das gesamte Web mittels *WebCorp* als Korpus untersucht werden, so ergeben sich aufgrund des hohen Datenvolumens und immer noch geringer Speicher- und Rechenleistung sehr lange Wartezeiten, da alle Berechnungen in Echtzeit durchgeführt werden müssen, um eventuelle Veränderungen in der Weblandschaft entsprechend wiedergeben zu können. Dies erklärt auch, warum im *WebCorp* nicht alle grammatikalischen Suchanfragen in Form von POS-Tags (Parts of Speech) durchführbar sind. Eine detaillierte Suche nach bestimmten Satzmustern ist, aufgrund der hier dargelegten technischen Einschränkungen, derzeitig nicht möglich. Ein aus linguistischer Sicht zentraler Einwand gegen *WebCorp* liegt in der Unbestimmbarkeit der Gesamtgröße des Korpus. Folgt man Lemnitzer und Zinsmeister (2006), so müssen empirische Grundstrukturen gesichert sein. Nur auf diese Weise können Korpora zu Forschungszwecken erstellt und zum Beleg herangezogen werden (vgl. Lemnitzer und Zinsmeister 2006: 14-40). Ferner kann bei dieser Form der Webanalyse nicht auf textkompositorische bzw. textklassifikatorische Charakteristika Rücksicht genommen werden, oder anders ausgedrückt: Nicht alles im Web Geschriebene ist auch im linguistischen Sinne Text. Das Web beinhaltet nur wenige für die linguistische Untersuchung bedeutende Metadaten wie z. B. Verfassungsdatum, Autor und sogar Sprache. Im Web existieren zahlreiche Seiten nebeneinander, jedoch unterscheiden sich diese Seiten und deren Informationsgehalt stark voneinander. Sprache verändert sich

und deshalb muss auch die diachrone Sprachbetrachtung bei Webseiteninhalten gegeben sein. Doch wer einmal versucht hat, auf jeder besuchten Webseite das Erstellungsdatum zu finden, wird schnell feststellen, dass diese Information nicht überall gegeben ist. Darüber hinaus unterscheiden sich Webseiten je nach Zielgruppe, Autoren und deren Kenntnissen in HTML. Im Gegensatz zu Printmedien gibt es, außer im akademischen Umfeld, kaum gestalterische und stilistische Richtlinien zum Verfassen und Gestalten von Webinhalten. Dies ist der Grund, warum Korpora aus Webdaten bisher nicht für empirisch-linguistische Forschungen herangezogen werden konnten (vgl. Kehoe und Gee 2007).

Um diesem Problem zu begegnen, wurde *WebCorp LSE* als Nachfolger des *WebCorp* erstellt. Die derzeit in *WebCorp LSE* verwendete Lösung sieht eine zweistufige Hybridversion vor.

Sind entsprechende Parameter, wie etwa ausreichender Speicherplatz und Rechenleistung, erfüllt, so kann das Web als Datenquelle für das Erstellen eines Korpus erschlossen werden.

Um die Daten zu filtern, wird aus bisherigen Anfragen und Ergebnissen ein zweites Offlinekorpus erstellt, das Eckdaten liefert und somit die Quantifizierung und Verbesserung der endgültigen Trefferanzeige begünstigt. Auf diese Weise kann, auf lange Sicht betrachtet, ein Abbild des Webs in Form eines durchsuchbaren Korpus geschaffen werden. Für die empirische Forschung und vor allem für statistische Untersuchung sollten Korpora eine bestimmbare Größe haben. Im „Synchronic English Web Corpus“ von *WebCorp LSE* liegen 467.713.650 Tokens vor, der „Diachronic English Web Corpus“ enthält 128.951.238 Wörter. Ferner müssen die Daten aufgrund der erwähnten Unterschiede in Komposition und Seitenaufbau aufbereitet werden.

Werden im *COCA* Magazine, Bücher und andere Printerzeugnisse aufbereitet, stammen die Daten in *WebCorp LSE* ausschließlich aus Webinhalten. Ein weiter Unterschied zwischen *COCA* und *WebCorp LSE* besteht deshalb auch in der Tiefe der Annotation. Bedingt durch die große Datenmenge, können in *WebCorp LSE* nur einige grundlegende syntaktische Inventarisierungen und Tokenisierungen vorgenommen werden. Die syntaktische Analyse beschränkt sich somit auf (semi-)automatische POS-Tagger. Grafisch und vereinfacht dargestellt ist *WebCorp LSE* folgendermaßen aufgebaut:

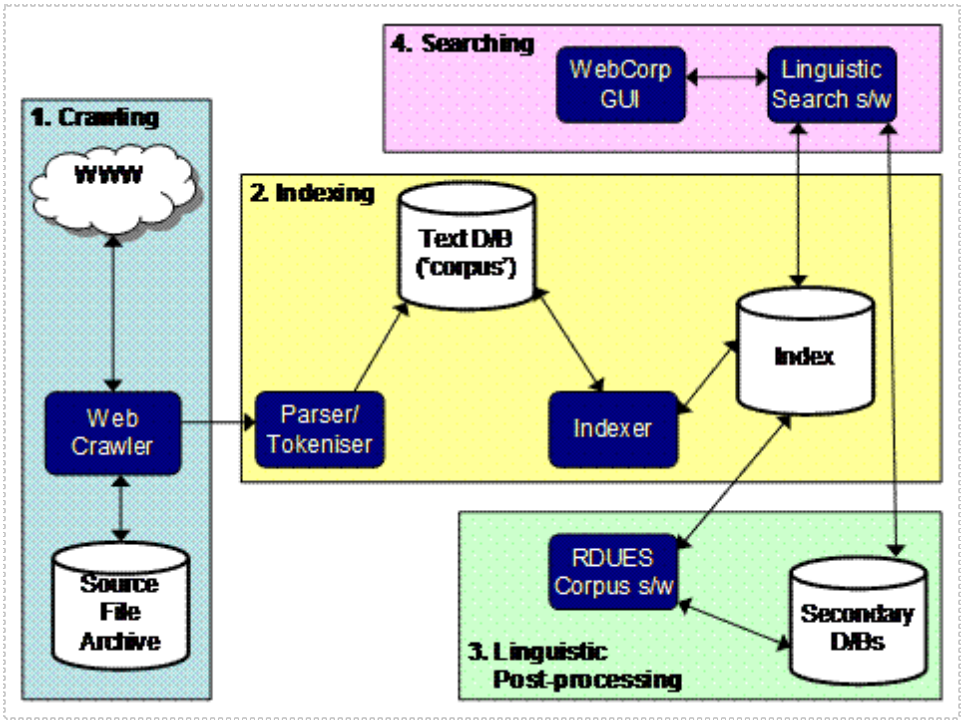

Abb. 1: Architektur und Funktionsweise von *WebCorp LSE* (Renouf 2009)

WebCorp LSE verfügt über Webcrawler, mit denen Inhalte aus dem Web als Texte heruntergeladen werden können. Ferner identifizieren und laden Webcrawler Links zu anderen Webseiten mit gleichem oder ähnlichem Inhalt herunter. Sind die Daten heruntergeladen, werden sie mittels computerlinguistischer Software (Parsern und Tokenisern) für den eigentlichen Output aufbereitet. Es gilt dabei jedoch zu beachten, dass *WebCorp LSE* nicht den Inhalt des gesamten Web speichert, da mittels der verwendeten Tools auch Texte und Inhalte von geringer Qualität im Vorfeld herausgefiltert werden. Unter Texten von geringer Qualität verstehen die Entwickler von *WebCorp LSE* Texte, die entweder zu kurze bzw. zu lange Satzstrukturen aufweisen. Auch die Länge einzelner Sätze spielt bei der Bewertung eine Rolle. Speziell entwickelte Programme filtern auch auf der Satzebene zu lange bzw. zu kurze Sätze heraus. Des Weiteren werden jene Webseiten automatisch aus dem Korpus entfernt, die hauptsächlich aus Hyperlinks und/oder Listen bestehen.[3]

[3] Gemäß der Erfahrung der Entwickler von *WebCorp LSE* dienen Webseiten mit übermäßig vielen Hyperlinks oder Listen nur dazu, Spam-Programme zu füttern. Für linguistische Untersuchungen sind diese Seiten somit nicht nutzbar.

Entsprechende Webseiten werden von der Software erkannt (*boilerplate detection*) und aus dem Korpus entfernt. Dies führt dazu, dass Webseiten, die nicht mindestens über einen Fließtext von 150 Wörtern verfügen, nicht in das Korpus aufgenommen werden. Bezogen auf das Kriterium der Satzlänge bedeutet dies, dass Texte, in denen die Hälfte der Sätze weniger als 50 Wörter aufweisen, ebenfalls nicht in das Korpus aufgenommen werden.[4]

Werden alle Kriterien berücksichtigt, ergibt sich ein Abbild der im Web verfügbaren Texte, die mit *WebCorp LSE* abgefragt werden können. Laut den Entwicklern ist das Ziel von *WebCorp LSE*:

> to build a 10 billion token web corpus over 2 years, consisting of:
>
> - a series of domain specific sub-corpora, updated monthly
> - newspaper sub-corpora, updated daily
> - a multi-terabyte 'mini web', updated monthly
>
> (Kehoe und Gee 2007)

Bei der Auswahl und Programmierung der Tools zur Aufbereitung der im Web gesammelten Texte muss neben den üblichen POS-Taggern auch auf die textkompositorische Natur des Webs eingegangen werden. Wie der Name *hypertext mark-up language* andeutet, werden mit HTML Webseiten programmiert und nicht nur Texte produziert. Somit ist ein Großteil der im HTML-Code verwendeten Begriffe rein gestalterisch und demnach semantisch arm. Ein Webcrawler erschließt und speichert jedoch den gesamten HTML-Code. Soll das Web als Korpus fungieren, so muss auch der linguistische Textbegriff für Texte aus dem Web überdacht werden (vgl. Ide 2002 und Kehoe und Gee 2007). Zusammenfassend betrachtet, stellten die Entwickler von *WebCorp LSE* folgende Richtlinien für die Untersuchung der Webseiten, deren Inhalt sowie deren Einbindung in die Korpusstruktur auf:

> A text should:
>
> - contain connected prose

[4] Informationen aus einer persönlichen Rücksprache mit Andrew Kehoe, Unit Director des Research and Development Unit for English Studies (RDUES) der Birmingham City University.

- be written in sentences, delimited by full stops

- contain paragraphs

- be complete, cohesive and interpretable with itself

(Kehoe und Gee 2007)

Um diese Kriterien zu erfüllen, programmierten die Entwickler eine Reihe von Tools, mit denen rein gestalterisches HTML (sogenannte *boilerplates*) während des Einsatzes von Webcrawlern erkannt und herausgefiltert werden.

In einem weiteren Arbeitsgang werden die HTML-Dateien (semi-)automatisch in korpustaugliche Formate wie etwa XML umgewandelt. Auch hierfür wurden eigens Tools entwickelt. Für detaillierte Informationen in Bezug auf Konzeption und Verfahrensweise sei an dieser Stelle an Kehoe und Gee 2007 verwiesen (vgl. auch *WebCorp LSE* „User Guide“).

Nach der Registrierung kann *WebCorp LSE* unter Eingabe von Benutzername und Passwort verwendet werden. Zu Beginn der Suche muss ein Referenzkorpus ausgewählt werden. Zu diesem Zweck können folgende Korpora ausgewählt werden: ‚Synchronic Web Corpus'; ‚Diachronic Web Corpus'; ‚Birmingham Blog Corpus', ‚Small French Newspaper'; ‚Anglo-Norman Correspondence Corpus'; ‚Charles Dickens Novels'; ‚Restoration Plays'; ‚Works of James Joyce'; ‚Works of Samuel Beckett'; ‚Works of Percy Bysshe Shelley'.

Für den Testdurchgang wurde das ‚Synchronic Web Corpus' ausgewählt. Durch Auswahl des Korpus gelangt der Benutzer zur eigentlichen Suchmaske.

Abb. 2: Die Benutzeroberfläche von *WebCorp LSE*

Unter ‚Query' wird der Suchbegriff eingegeben. Die Schaltfläche ‚Insert part of speech tag' bietet verschiedene Optionen bzw. voreingestellte Operatoren zur Berücksichtigung von POS-Tagsets. Dies dient vor allem der Suchoptimierung. Mit der Schaltfläche ‚sub-corpus' kann das Sachgebiet eingrenzt werden. Dem Benutzer stehen 15 Sachgebiete zur Verfügung: ‚Mini-web sample'; ‚Arts'; ‚Business'; ‚Computers'; ‚Games'; ‚Health'; ‚Home'; ‚Kids and Teens'; ‚News'; ‚Recreation'; ‚Reference'; ‚Science'; ‚Shopping'; ‚Society'; ‚Sport'. Mit der darunterliegenden Schaltfläche ‚case sensitive' kann die Unterscheidung zwischen Groß- und Kleinschreibung ein-/ausgeschaltet werden. Für die folgende Untersuchung wurde kein ‚sub-corpus' ausgewählt, um die Leistungsfähigkeit der gesamten *WebCorp Linguist's Search Engine* zu berücksichtigen. Des Weiteren kann unter ‚sentence position' die Position des Suchwortes im Satz festgelegt werden. Hier stehen die Optionen ‚any', ‚within a single sentence', ‚sentence initial' und ‚sentence final' zur Verfügung. Auch diese Optionen dienen

der Suchoptimierung. Mit der Option ‚word filter' können Wörter festgelegt werden, die in der Suchanfrage vorkommen müssen bzw. ausgeschlossen werden sollen. Die Reichweite dieser Suchoption kann sich auf das gesamte im Korpus verfügbare Dokument, einzelne Sätze oder die Position des Wortes im Satz beschränken. Es können auch mehrere Wörter innerhalb der Suchanfrage ausgewiesen werden. Die Entfernung links oder rechts zum Suchwort kann manuell zwischen Position 1-30 festgelegt werden. Soll ein Wort nicht vorkommen, muss dieses durch Minuszeichen vor dem entsprechenden Wort markiert sein. Wird beispielsweise nach dem Wort ‚plant' gesucht, kann die Suche durch die Auswahl von ‚nuclear-flower' unter Verwendung des Operators ‚document' eingegrenzt werden. Dies führt dazu, dass nur Treffer, die das Wort ‚nuclear' und nicht ‚flower' beinhalten, angezeigt werden. Zu jeder der einzelnen Optionen steht eine kontextbezogene Hilfefunktion zur Verfügung.

3 *WebCorp LSE* vs. *COCA*

Im folgenden Testdurchgang wurden die in der Einleitung erwähnten Termini in beide Korpora eingegeben. Die Suche beschränkte sich auf Kollokationen des Typs [VP]+[NP] bzw. [NP]+[VP],[5] da diese bei der Übersetzung der Fachtexte aus dem Themenbereich der Zementherstellung die meisten Probleme bereiteten.

Die erste Spalte der Tabelle spiegelt die Zahl der Gesamttreffer wider, in der zweiten Spalte werden die untersuchten Verbal-/Nominalphrasen – Kollokationen aufgelistet. Lieferte die Suche mit *WebCorp LSE* mehrere Treffer, wurde mittels des Suchoperators {V*} nach möglichen Kollokationen in Vor- bzw. Nachstellung gesucht. Die auf diese Weise erzielten Treffer werden ebenfalls aufgelistet. Doppelte Vorkommen wurden nicht herausgefiltert. Sofern vorhanden, sind die Ergebnisse der Suche nach Vor- bzw. Nachstellung durch die Kollokationsmuster gekennzeichnet. Es folgen die Ergebnisse in tabellarischer Form:

[5] Siehe Hausmann 1999 für eine detaillierte Besprechung der unterschiedlichen Kollokationstypen.

Term	*LSE*	[VP]+[NP]	[NP]+[VP]	*COCA*	[VP]+[NP]	[NP]+[VP]
initial setting time	0	0	0	0	0	0
flexural strength	0	0	0	0	0	0
rotary kiln	7	0	0	5	0	1
compressive strength	13	0	0	29	1	2
aggregate	1721	539	395	1836	119	59
limestone marl	0	0	0	1	0	0
calcining zone	0	0	0	0	0	0
wet process	35	0	10	6	0	0
raw meal	8	0	4	1	0	0
pore formation	21	0	0	0	0	0
transport cement	2	0	0	1	0	0

Bei Betrachtung der Ergebnisse fällt zunächst auf, dass beide Korpora wenige, bis keine Treffer bei Mehrworttermini erzielten. Die Ausnahmen lagen hier vor allem bei der Suche mit *WebCorp LSE* bei den Begriffen ‚*wet process*' und ‚*pore formation'*. Die Treffer sind jedoch nicht dem Bereich ‚Zementarten und Zementherstellung', sondern vorwiegend anderen Fachgebieten zuzuordnen. Folglich muss die Disambiguierungsleistung vom Benutzer bzw. dem Übersetzer erbracht werden. Der Zeitaufwand für diese manuelle Analyse der Type-Token-Liste steigt mit zunehmender Trefferzahl. Dieses Phänomen wird durch die hohe Anzahl der Treffer beim Suchbegriff ‚*aggregate'* verdeutlicht.

Die fachspezifische Bedeutung konnte nur in begrenztem Maße aus den Korpora extrahiert werden. Dies geschah zum einen durch die genaue Spezifikation der Wortart, also

‚*aggregate* [NN]' (nur Instanzen von ‚*aggregate*' als Nomen) und zum anderen mittels einer Durchsicht der Liste.

Beim Testdurchgang in *COCA* fiel positiv auf, dass durch die Angabe der Quelle in der Suchmaske die Einträge schneller einem Sachgebiet zugeordnet werden konnten, bei *WebCorp LSE* hingegen musste hierzu ein Fenster aufgerufen werden. Eine manuell gefilterte Liste aus *WebCorp LSE*, die dem Fachgebiet Zementarten und Zementherstellung entspricht, befindet sich im Anhang. Bei der Durchsicht tritt die Kollokation *aggreagte* + *should* + *be* zwölf Mal auf, gefolgt von *aggregate* + *is* mit acht Treffern. Fachkollokationen ließen sich mit *aggregate* + *saturated* und *aggregate* + *grading* nachweisen (vgl. Anhang).

Die geringe Zahl verwertbarer Kollokationen für ‚*limestone marl*' ist durch die fachsprachliche Verwendung der Begriffe durch Sachkundige und Experten begründet. Die Treffer für ‚*raw meal*' sind ebenfalls nicht dem Sachgebiet der Zementarten, sondern dem Agrarsektor bzw. der Futtermittelindustrie zuzuordnen. In diesen Industriezweigen wird der Begriff für aufbereitetes Getreide verwendet. Die Angabe der Treffer bei ‚*transport cement*' mag auf den ersten Blick verwundern, da ‚*transport*' und ‚*cement*' eine häufige Kollokation ist. Jedoch lieferten beide Korpora in diesem Fall keine Treffer, da ‚*transport*' als Substantiv vordefiniert wurde. Bei der Eingabe von ‚*transport cement*' werden jedoch nur solche Treffer des Musters [NP]+[NP] aufgelistet, da bei der Suche die deutsche Entsprechung als Hauptwort gefunden werden sollte.

4 Ausblick

Zusammenfassend betrachtet ist festzustellen, dass *WebCorp LSE* geringfügig mehr Treffer bei Termini liefert, die eindeutig dem fachsprachlichen Bereich zuzuordnen sind. Dieses Ergebnis ist jedoch nicht als Empfehlung zu verstehen, sich bei Fachtexten auf die Suche mit *WebCorp LSE* zu beschränken, da auch hier die Trefferquote nicht hoch genug ist. Die eingangs aufgestellte Hypothese, dass fachsprachliche Termini und ihre Kollokationen in den verfügbaren Korpora in diesem Test unzureichend abgebildet sind, scheint somit bestätigt. Beide Korpora unterscheiden nicht zwischen allgemeinsprachlichem und fachsprachlichem

Gebrauch. Ferner zeigen beide Korpora deutliche Schwächen bei Anfragen zu Mehrworttermini, hier vor allem bei dreigliedrigen Komposita.

Ein weiteres Problem besteht darin, dass der Benutzer, wie z. B. bei der Anfrage zu ‚*transport cement*' die Wortarten genau festlegen muss, um die gewünschte Kollokation zu finden.

Gegenwärtig ist keines der beiden Korpora für die fachterminologische Kollokationsrecherche geeignet. Die Benutzeroberfläche von *COCA* bietet jedoch mehr Funktionen und differenzierte Optionen zur Filterung von Suchanfragen an.

Längerfristig betrachtet wäre die Entwicklung einer umfangreichen und leistungsstarken Suchmaschine, die einen direkten Zugriff auf den im Web verfügbaren Inhalt ermöglicht, erstrebenswert. Vor allem die Disambiguierung von eingegeben Suchbegriffen ist zentral für den Übersetzungsprozess. So kann zu Beginn der Suche festgelegt werden, dass die allgemeinsprachliche Verwendung des Suchbegriffes ausgeschlossen wird. Vom technischen Standpunkt betrachtet müsste die Suchmaschine also über ein internes Wörterbuch verfügen oder die eingegeben Suchbegriffe direkt über webbasierte Wörterbücher semantisch aufschlüsseln, damit der Benutzer die gewünschte Bedeutung aus einer Liste auswählen kann. Für die englische Sprache wäre es beispielsweise möglich, die Informationen aus *The Free Dictionary* zu entnehmen. Die Suchmaschine *DuckDuckGo* verfährt bei der Eingabe von Begriffen auf diese Weise. Auch eine Anwendung innerhalb der Korpusbenutzeroberfläche wäre so zu erreichen.

Ebenfalls müssten Testverfahren für die übersetzungsbezogene Verwendung von Korpora erstellt werden. Dies schließt unter anderem auch ein, welche Wortlisten aus welchen Bereichen für welches Korpus verwendet werden sollten. Auch in der Computerlinguistik gibt es noch kein standardisiertes Format für die Erstellung von Korpora bzw. entsprechender Testwerkzeuge. Aufgrund der unterschiedlichen Leistungserwartungen an Korpora wäre ein Austausch zwischen Übersetzern und (Computer-)Linguisten wünschenswert. Gegenwärtig forscht Burghardt über die Usability von Annotationswerkzeugen.

Schließlich könnte im Rahmen eines Gemeinschaftsprojektes ein translatorisches Gesamtkorpus erstellt werden. Die Herausforderungen bei der Erstellung eines solchen Korpus liegen in der Darstellung und Annotation mehrsprachiger Daten sowie der Disambiguierung.

Versuche dieser Art wurden erfolgreich für juristische Fachtexte im *European Parliament Parallel Corpus* bzw. für Patente im Webauftritt des Europäischen Patentamtes durchgeführt. Auf Grundlage dieser Projekte ließe sich ein erstes translatorisches Gesamtkorpus für mehrere Sprachen realisieren.

Eine webbasierte Lösung in Form einer Extraktion von Übersetzungen ist momentan nicht umsetzbar. Ganz gleich wie gut die Qualität der Übersetzungen auch ist, sie muss dennoch, vor Aufnahme in ein Korpus, manuell überprüft werden. Dies führt zu einem erhöhten Arbeitsaufwand, besonders bei mehrsprachigen Korpora.

Ferner führt die Mehrsprachigkeit der Einträge zu einer erhöhten Datenmenge, die momentan noch nicht abzuschätzen ist. Die Verwendung von Texten und Übersetzungen, die innerhalb der europäischen Union produziert werden, als Grundlage für ein solches Gesamtkorpus, entspricht nicht unbedingt dem gewünschten Abbild des Sprachgebrauchs. Dies liegt in erster Linie darin begründet, dass auch die EU-Institutionen eigene fachsprachliche Termini verwenden. Für das Sprachenpaar Deutsch-Englisch ließe sich ein Referenzkorpus aus den Webseiten *Linguee* oder *MyMemory* erstellen. Doch auch hier muss ein qualitativer Filterprozess zur Bewertung der Termini bzw. Übersetzungen vorgeschaltet werden.

5 Bibliografie

Benson, Morton, Evelyn Benson und Robert Ilson (eds.) (1999). *Student's Dictionary of Collocations*. Berlin: Cornelsen.

Bilić, Viktorija, Martha Connelly und Joachim Kornelius (2009). *Wissensrecherche als kooperatives Handeln*. Trier: Wissenschaftlicher Verlag Trier.

Burghardt, Manuel (2012). *Usability von Annotationswerkzeugen*. Regensburg.

Fluck, Hans-Rüdiger (1996). *Fachsprachen: Einführung und Bibliographie*. 5. Aufl. Tübingen [u. a.]: Francke.

Halkiopoulou, Sirmula (2006). *Syntagmatische Semantik im Kontext der fachsprachlichen Lokalisierung*. Heidelberger Studien zur Übersetzungswissenschaft 6. Trier: Wissenschaftlicher Verlag Trier.

Hausmann, Franz Josef (1999). „Praktische Einführung in den Gebrauch des *Student's Dictionary of Collocations"*. *Student's Dictionary of Collocations*. Morton Benson, Evelyn Benson und Robert Ilson (eds.). Berlin: Cornelsen. iv-xv.

Heid, Ulrich (2011). „Zur Extraktion von Fachwortschatz aus Texten". Vortrag am Seminar für Übersetzen und Dolmetschen der Ruprecht-Karls-Universität Heidelberg am 21. Januar 2011. http://www.uni-heidelberg.de/md/sued/seminar/abteilungen/italienisch/heid_folien_vortrag.pdf.

Ide, Nancy, Randi Reppen und Keith Suderman (2002). „The American National Corpus: More Than the Web can Provide". Poughkeepsie, New York. http://www.cs.vassar.edu/~ide/papers/anc-lrec02.pdf.

Kehoe, Andrew und Matt Gee (2007). „New corpora from the web: making text more ‚text-like'". *Varieng – Studies in Variation, Contacts and Change in English* 2/2007. http://www.helsinki.fi/varieng/journal/volumes/02/kehoe_gee/.

Lemnitzer, Lothar und Heike Zinsmeister (2006). *Korpuslinguistik: Eine Einführung*. narr Studienbücher. Tübingen: Narr.

McIntosh, Colin. (Hg.) (2009). *Oxford Collocations Dictionary for students of English*. Oxford: Oxford University Press.

Renouf, Antoinette (2009). „Corpus Linguistics beyond Google: the WebCorp Linguist's Search Engine". Digital Studies / Le champ numérique 1:1. http://www.digitalstudies.org/ojs/index.php/digital_studies/article/view/147.

Reske, Christian (2012). *Grau in Grau? Zementarten und Zementherstellung im Kontext der computergestützten Fachübersetzung*. Lighthouse Unlimited 160. Trier: Wissenschaftlicher Verlag Trier.

6 Internetquellen

Bing. http://www.bing.com/.

BootCaT – Simple Utilities to Bootstrap Corpora And Terms from the Web. Marco Baroni und Silvia Bernardini. http://bootcat.sslmit.unibo.it/?section=home.

British National Corpus (*BNC*). University of Oxford. http://www.natcorp.ox.ac.uk/.

Corpus of Contemporary American English (*COCA*). Mark Davies. Brigham Young University. http://www.americancorpus.org/.

Corpus of Contemporary American English (*COCA*). Mark Davies. „*Where should I start*?". Brigham Young University. http://corpus.byu.edu/coca/.

COSMAS II. COSMAS (Corpus Search, Management and Analysis System). Institut für Deutsche Sprache, Mannheim. http://www.ids-mannheim.de/cosmas2/uebersicht.html.

Deutscher Wortschatz – Portal. *Wortschatz Universität Leipzig*. Universität Leipzig, Institut für Informatik, Abteilung Sprachverarbeitung. http://wortschatz.uni-leipzig.de/.

DuckDuckGo. http://duckduckgo.com/.

DWDS – Digitales Wörterbuch der deutschen Sprache. Berlin-Brandenburgische Akademie der Wissenschaften. http://www.dwds.de/.

DWDS – *Digitales Wörterbuch der deutschen Sprache*. Berlin-Brandenburgische Akademie der Wissenschaften. „Spezialkkorpora". http://www.dwds.de/resource/spezialkorpora/.

DWDS – *Digitales Wörterbuch der deutschen Sprache*. Berlin-Brandenburgische Akademie der Wissenschaften. „Zeitungskorpora". http://www.dwds.de/resource/zeitungskorpora/.

EPO. European Patent Office / Europäisches Patentamt. „EPO - Espacenet“. http://www.epo.org/searching/free/espacenet.html.

European Parliament Proceedings Parallel Corpus 1996-2011. Philipp Koehn. http://www.statmt.org/europarl/.

Google. http://www.google.com/.

Linguee. http://www.linguee.de/.

MyMemory. Maschinenübersetzungen vs. Humanübersetzungen. http://mymemory.translated.net/.

The Free Dictionary. Dictionary, Encyclopedia and Thesaurus. Farlex Inc. http://www.thefreedictionary.com/.

The Linguist's Search Engine. http://lse.umiacs.umd.edu/.

WebCorp. Research and Development Unit for English Studies (RDUES) in the School of English at Birmingham City University. http://www.webcorp.org.uk/live/index.jsp.

WebCorp Linguist's Search Engine. Research and Development Unit for English Studies (RDUES) in the School of English at Birmingham City University. http://wse1.webcorp.org.uk/.

WebCorp Linguist's Search Engine. „Benutzeroberfläche“. Research and Development Unit for English Studies (RDUES) in the School of English at Birmingham City University. http://wse1.webcorp.org.uk/cgi-bin/SYN/index.cgi.

WebCorp Linguist's Search Engine. „Diachronic English Web Corpus“. Research and Development Unit for English Studies (RDUES) in the School of English at Birmingham City University. http://wse1.webcorp.org.uk/dia.html.

WebCorp Linguist's Search Engine. „Synchronic English Web Corpus“. Research and Development Unit for English Studies (RDUES) in the School of English at Birmingham City University. http://wse1.webcorp.org.uk/syn.html.

WebCorp Linguist's Search Engine. „User Guide". Research and Development Unit for English Studies (RDUES) in the School of English at Birmingham City University. http://wse1.webcorp.org.uk/guide/.

7 Anhang

Type-Token-Liste zum Begriff ‚*aggregate*' in *WebCorp LSE*. Die entsprechenden Kollokationen wurden manuell dem Sachgebiet der Zementarten und Zementherstellung zugeordnet.

bricks, and in the **aggregate used** in concrete. The concentration

proportions shall be based on the batch weight of each **aggregate saturated**, surface-dry weight plus the weight of surface moisture it contains

The quantity of each type of filter or bedding **aggregate delivered** and placed within the specified limits is computed to the

percent. 3) Other similar grout mixes that incorporate small coarse **aggregate may** be used but must be approved by the Engineer. 4)

4 inch thick of 3/8" to 3/4" crushed **aggregate is** placed on the bottom of the excavation. KEYSTONE AUDIO VISUAL

fine and coarse particles. The gradation of coarse and fine **aggregate is** very important. Very fine sand is not recommended for concrete

The surface texture is very rough. Therefore such fine **aggregate is** not suitable for concreting and other construction work. The Plasticy

of water. It may be used with any type of **aggregate including** light weight and porous material. The test is not affected

areas up to 5 acres. An upstream layer of smaller **aggregate may** be used for filtering. Rock can be placed by hand

subsoil from aggregate within a subsurface drain Separating subsoil from **aggregate placed** at the soil surface Stabilization of soil surface during temporary

of large rocks, chinking with smaller rocks and **aggregate, filling** with grout, surface finishing, and curing. Machined Riprap:

:VARIABLE (SEE FIGURE 2). WET STORAGE GEOTEXTILE Coarse **aggregate should** be TDOT 3 357, or 5. TDOT CLASS

Silt Fence SF for installation requirements. Clean coarse **aggregate should** be placed outside the box, all around the

a depth of 2 to 4 inches. Coarse **aggregate should** be TDOT 3 357, or 5. If the

it no longer adequately performs its function, the **aggregate should** be pulled away from the structure, cleaned, and

openings to hold gravel in place. Clean coarse **aggregate should** be placed up to 2 inches below the

flatter and smoothed to an even grade. Coarse **aggregate should** be TDOT 3 357, or 5. If the

with inch openings should be used. Clean coarse **aggregate should** be placed over the entire inlet structure, to

total depth of at least 12 inches. The **aggregate should** extend beyond the inlet structure at least 18

at least 18 inches on all sides. Coarse **aggregate should** be TDOT 3 357, or 5. Sediment should

beyond the upstream and downstream toe of the **aggregate placed** around the culvert. 5. The culvert(s) should

abrasiveness on equipment drying time bond of dry **aggregate dried** whiteness crack resistance coverage SF/lb spray(1)

sparser aggregate surface may also result. Also, if **aggregate is** accidentally brushed off, a lighter colored surface may

texture to provide a uniform texture appearance. Vary **aggregate grading**, aggregate proportion, number of passes over the surface,

gun or machine application equipment without catalyst. Vary **aggregate grading**, number of passes over the surface, air pressure

values and reduced weight. For sand-float finishes, the **aggregate should** be a fine silica sand. All aggregates should

with metal lath as the plaster base, perlite **aggregate is** not recommended for use in the basecoat plaster,

that will trap all particles larger than the **aggregate being** sprayed. The basecoat must be free of ridges

of sand for scratch and brown coats. Lightweight **aggregate should** not be used in replastering when using metal

accelerator alum catalyst as an accelerator when limestone **aggregate is** used. Remedy: None. Dispose of batch. Prevention: Use

quick-set gauging plaster as an accelerator when limestone **aggregate is** used, or use sand aggregate. Also, avoid entraining

to normal range. c. Cause: Excessive use of **aggregate. Remedy**: Patch affected area. Prevention: Use proper proportions of

Cause: Too much aggregate or fine, poorly graded **aggregate. Remedy**: No remedy; remove and replaster. Prevention: Use properly

a." Prevention: Reduce mixing time. c. Cause: Poor **aggregate. Remedy**: See above in "a." Prevention: Use clean, properly

sustainable. ready-mixed plaster A calcined gypsum plaster with **aggregate added** during manufacture. A powder product that requires the

stucco (1) A mixture of Portland cement and **aggregate designed** for use on exterior or interior surfaces exposed

steel deck roadway that has a factory-installed bonded **aggregate wearing** surface. State-of-the-Art Advanced Traffic Management System (ATMS Operates

pot holes, clearing/ repairing drainage, lifting debris, loading **aggregate required** for construction or repair, digging trenches for culverts

order to produce that particular aggregate. If the **aggregate be** of pebbles, and we call it two, the

the desire to keep construction costs down. Concrete **aggregate collected** from demolition sites is put through a crushing

strength and constitutive property behavior of a fine **aggregate cemented** material (FACM). The FACM was designed to have

programs. Id. An average of 11 tons of **aggregate is** required annually for each and every resident in

Pattern [VP] + [NP]

importance of taking geological aspects into account when **planning aggregate** extraction and use in the construction industry Geological

and other wetlands and in streambeds). NSSGA's members **excavate aggregate** materials and create large depressions that may fill

industrial look. The anchors are provided in an **exposed aggregate** finish for an aesthetically pleasing result. Sliding Gate.

Market. Payback of Mining Activities Within Entropia Universe. **Measuring aggregate** production in a virtual economy using log data.

influence of the proprietary geogrid performance properties with **increasing aggregate** base course thickness. Therefore, it appears to be

cycles are considered, like in Table 2, the **unreinforced aggregate** thicknesses are quite different. Although geotextile aggregate thickness

this environmental tax, intended to discourage quarrying and **encourage aggregate** recycling, was unlikely to operate satisfactorily in Northern

1960 by a group of conscientious and environmentally **concerned aggregate** producers. The Association's goal is to protect and

T21N - Translation in Transition

T21N offers a cutting-edge electronic publishing venue, created by experts for both young talent and established researchers from the worlds of translation and interpreting.

T21N provides a stage for emerging ideas and new academic talent to present their ideas in a digital reading site, where speed and ease meet enjoyment.

T21N is exclusively published online at http://www.t21n.com.

Articles in compliance with our style sheet may be submitted at any time and will be published at short notice.

T21N editors research and teach at the Institute of Translation and Interpreting at the University of Heidelberg in Germany.