Daniel Bunčić (Tübingen)

Film-Untertitel als Quelle eines quasi-mündlichen Parallelkorpus?*

1. Einleitung

Dieser Beitrag befasst sich mit der Möglichkeit, die auf vielen handelsüblichen DVDs enthaltenen Untertitelspuren in verschiedenen Sprachen zu nutzen, um ansonsten meist ausschließlich auf schriftlichen Texten basierende Parallelkorpora auszubalancieren. Die zentrale Frage, die sich hierbei stellt, ist, inwiefern diese ja in schriftlicher Form vorliegenden (und daher besonders leicht in ein Korpus zu integrierenden), aber gesprochene Rede wiedergebenden Untertitel ‚mündlicher' sind als andere schriftliche Texte.
Um diese Frage beantworten zu können (9.), wird die Entstehung der mehrsprachigen Untertitel in drei Schritte zergliedert, die einzeln zu analysieren sind: die *Fiktionalisierung* von Dialog im Film gegenüber realem, nicht-fiktionalem Dialog (8.), den *Medienwechsel* bei der schriftlichen Wiedergabe des Filmdialogs in Untertiteln (6.) und die *Übersetzung* des Untertitels der Originalsprache in eine Zielsprache (7.). Aus praktischen Gründen wird dabei die Fiktionalisierung zuletzt behandelt, da es einfacher ist, fertige Untertitel mit Transkripten realer Dialoge zu vergleichen als mündliche Filmdialoge mit mündlichen Realdialogen. Vor dieser Analyse ist jedoch zu klären, welchen Zweck Parallelkorpora eigentlich haben (2.), was in diesem Zusammenhang unter Mündlichkeit zu verstehen ist (3.) und welche Ansätze zu mündlichen Korpora es bisher gibt (4.). Zudem werde ich darstellen, wie die Untertitelspuren aus einer DVD ausgelesen werden können und welche Probleme sich dabei ergeben (5.).

---

* Diese Schriftfassung beruht auf einem Vortrag, der schon 2009 beim 18. JungslavistInnen-Treffen in Regensburg sowie in ähnlicher Form beim 10. Deutschen Slavistentag in Tübingen vorgetragen wurde. Neben den Diskussionen nach den beiden Vorträgen sind in diese Fassung auch Informationen der Untertitel-Übersetzerin Beatrix Kersten eingegangen, die am 20. April 2010 einen Vortrag an der Universität Tübingen gehalten hat und der ich für den anschließenden Austausch in besonderem Maße zu Dank verpflichtet bin. Dank schulde ich ebenfalls Ruprecht von Waldenfels und Roland Meyer für ihre Hilfe bei der Integration der Untertitel in das *ParaSol*-Korpus.

2. Parallelkorpora

In den letzten Jahren ist ein wachsendes Interesse an mehrsprachigen Korpora zu verzeichnen. Im Bereich slavischer Sprachen wurden bisher u.a. folgende für die wissenschaftliche Öffentlichkeit zugängliche[1] Parallelkorpora entwickelt:

- *ParaSol (A Parallel Corpus of Slavic and other languages*, früher *Regensburg Parallel Corpus, RPC)*, das derzeit[2] Texte in insgesamt 12 slavischen und 12 nichtslavischen Sprachen enthält (http://www-korpus.uni-r.de/ParaSol/, 14 Mio. slavische Tokens)
- das in das Tschechische Nationalkorpus integrierte Projekt *InterCorp* mit Texten auf Tschechisch und in 7 weiteren slavischen und 15 nichtslavischen Sprachen (http://www.korpus.cz/intercorp/, 66 Mio. slavische Tokens)
- das Modul *KoParT (Korpus parallel'nych tekstov)* innerhalb des Russischen Nationalkorpus, derzeit mit englisch-, deutsch- und ukrainisch-russischen Bitexten (http://ruscorpora.ru/search-para.html, 9 Mio. Tokens, laut Dobrovol'skij, Kretov/Šarov 2005, 263 f. in der Pilotphase 2,7 Mio. russische Tokens)
- das *GRALIS*-Textkorpus, das bisher nur bosnische, kroatische und serbische Texte enthält (http://www-gewi.uni-graz.at/cocoon/gralis/; zusammen 2 Mio. Tokens)
- *Europarl*, ein Korpus aus Transkripten von Debatten des Europaparlaments, das u. a. die 5 slavischen Sprachen der EU enthält (http://www.statmt.org/europarl/, 55 Mio. slavische Tokens)
- *OPUS*, eine gemeinfreie Sammlung mehrerer großer Parallelkorpora (http://opus.lingfil.uu.se/), z. B. *SETimes* aus Texten der *Southeast European Times* auf Englisch und in 8 Balkansprachen, darunter 4 slavische (149 Mio. slavische Tokens)

Schon 1971 gab es in Zagreb ein durch Übersetzung des halben Brown-Korpus entstandenes serbokroatisch-englisches Parallelkorpus (mit je 500 000 Tokens), das inzwischen aber nicht mehr zugänglich ist (vgl. Tadić 2000, 523). Die ersten Ansätze zu mehrsprachigen elektronischen Korpora aus den 90er Jahren boten im Rahmen der *Trans-European Language Resources Infrastructure (TELRI)* Platons *Politeia* in 5 slavischen und 5 nichtslavischen Sprachen (http://nl.ijs.si/telri/Republic/) sowie im Rahmen von *Multext-East* Orwells *1984* zunächst in 3 slavischen und 3 nicht-slavischen, in der inzwischen 4. Ausgabe aber nun in 11

[1] Das von Orechov (2009) beschriebene Korpus der Übersetzungen des Igorlieds ist im Moment anscheinend aus technischen Gründen nicht zugänglich.

[2] Alle Angaben zu Korpora (einschließlich der weiter unten wiedergegebenen Suchergebnisse) und Internetadressen sind auf dem Stand vom März 2011.

slavischen und 6 nichtslavischen Sprachen (http://nl.ijs.si/ME/). Sie enthielten nur je gut 100 000 Tokens pro Sprache und sind daher quantitativ weit überholt – auch wenn etwa die resianische Version von *1984* ein gewisses ‚Alleinstellungsmerkmal' des *Multext-East*-Korpus darstellt.

Solche Parallelkorpora erfüllen andere Funktionen als einsprachige Korpora. Wer etwa einen wasserdichten Beleg für ein sprachliches Phänomen sucht, ist gut beraten, dazu auch einsprachige Korpora und/oder Probandenbefragungen zu benutzen, da die ‚Beweiskraft' von übersetzten Texten für solche Zwecke eingeschränkt ist: "Translation […] reeks of non-authentic language, meaning: one can never be sure whether a given phenomenon that is attested in a translation would ever have been produced in the same way in an original text of the same language" (Stolz 2007, 102). Hinzu kommt, dass Parallelkorpora im Allgemeinen eine geringere Bandbreite an Textsorten als einsprachige Korpora enthalten (dazu unten mehr). Andererseits hat man bei einsprachigen Korpora das Problem der Vergleichbarkeit. Parallelisierte Korpora hingegen erlauben die Untersuchung konkret-extensional statt nur abstrakt-intensional definierter Kategorien und können bei typologischen Vergleichen helfen, Infor-mationen über bisher wenig beachtete Phänomene zu finden, auf die z.B. Grammatiken nicht eingehen (ausführlich zu den Vor- und Nachteilen von Parallelkorpora vgl. Stolz 2007 und Wälchli 2007). Vor allem aber haben sich Parallelkorpora in den letzten Jahren als überaus nützlich für das Übersetzungswesen (sowohl als Hilfe für Über-setzerInnen als auch als Grundlage automatischer Übersetzungsprogramme) und für die Lexikographie (gedruckte wie elektronische Wörterbücher) erwiesen (vgl. Simeon 2002).

3. Mündlichkeit

Nach Koch/Oesterreicher (1985; 2007) besteht ein wichtiger Unterschied zwischen *Medium* und *Konzeption*. In Bezug auf das Medium ist ein mündlicher *(phonischer)* Text trivialerweise einer, der aus Lauten besteht, durch Schallwellen übertragen und mit den Ohren wahrgenommen wird, während ein schriftlicher *(graphischer)* Text in Schriftzeichen festgehalten, durch Lichtwellen übertragen und mit den Augen wahrgenommen wird. Viel wichtiger für die sprachliche Form eines Textes ist jedoch der konzeptionelle Aspekt. So gibt es konzeptionell *geschriebene* Texte, die aber phonisch vorgetragen werden, wie etwa den wissenschaftlichen Vortrag (Koch/Oesterreicher 2007, 349). Die augenfälligsten Beispiele für graphisch übertragene, aber konzeptionell *gesprochene* Texte sind wohl Chat und SMS. Auf

solche graphisch repräsentierte Mündlichkeit bezieht sich der Ausdruck *quasi-mündlich* im Titel dieses Beitrags.

Für die sprachliche Variation entlang des Kontinuums zwischen konzeptionell gesprochenen und konzeptionell geschriebenen Texten haben Koch/Oesterreicher (1985, 23) die Ausdrücke „Sprache der Nähe“ und „Sprache der Distanz“ geprägt. Koch (1999) legt ausführlich dar, dass dieser Unterschied nicht unter die diaphasische Dimension sprachlicher Variation subsumiert werden kann. Allerdings haben die Autoren für die Nähe-Distanz-Opposition kein in das Coseriu’sche Schema passendes Adjektiv geprägt. Kabatek (2000, 313–314) verweist auf eine italienische Tradition, die für diesen Sachverhalt den Terminus *diamesisch* (zu gr. *μέσον* ‘Mitte’, das genau lat. *medium* entspricht) benutzt.

Wenn im Folgenden von Mündlichkeit die Rede ist, so geht es nicht um die mediale Dimension,[3] sondern um das Problem, dass Parallelkorpora in diachronischer, diatopischer und diaphasischer Hinsicht verschiedenartige Texte enthalten können, diamesisch jedoch bisher fast ausschließlich die „Sprache der Distanz“ repräsentieren.

4. Mündliche Korpora

In der Anfangszeit der Korpuslinguistik bestanden Korpora ausschließlich aus geschriebenen (und oft mehrheitlich literarischen) Texten, weil diese erheblich einfacher in ein Korpus zu integrieren sind als gesprochene Texte. Inzwischen aber enthalten die großen einsprachigen Korpora oft auch mündliche Subkorpora, so etwa das Russische Nationalkorpus in seinem *Korpus ustnoj reči* (http://www.ruscorpora.ru/search-spoken.html)[4] oder das Tschechische Nationalkorpus, das fünf

[3] Für phonetische Untersuchungen, für die tatsächlich phonisch, z.B. als Klangdatei, vorliegende Sprachaufzeichnungen von Interesse sind, gibt es inzwischen erste Audiokorpora. Beispielsweise enthält das Russische Nationalkorpus ein „multimediales“ Korpus, bei dem das Transkript mit einer entsprechenden Tondatei verlinkt ist (http://www.ruscorpora.ru/search-murco.html; vgl. z. B. Grišina/Savčuk 2008), und auch die in Fn. 5 erwähnten Kurzdialoge des *Multext-East*-Korpus sind in einigen Sprachen als Tondateien erhältlich. Eine solche Möglichkeit besteht theoretisch zum Teil auch für Untertitel-Korpora (jedenfalls für die Originalsprache und evtl. bei ebenfalls vorhandenen Synchronfassungen, die allerdings textlich meist von den Untertiteln abweichen); dies kann hier jedoch nicht weiter verfolgt werden.

[4] Von den 7,5 Mio. Tokens, die dieses Subkorpus 2008 enthielt (inzwischen sind es 9,5 Mio.), entfiel jedoch lediglich ein Zehntel auf Aufzeichnungen nichtöffentlicher Gespräche, der

mündliche Subkorpora mit je einer halben bis einer Million Tokens enthält, darunter ein „soziolinguistisch ausgewogenes Korpus informell gesprochenen Tschechischs" („socio-lingvisticky vyvážený korpus neformální mluvené češtiny", http://ucnk.ff.cuni.cz/struktura.php).

Die angeführten Parallelkorpora hingegen bestehen bisher auch weiterhin fast ausschließlich aus schriftlichen Texten (was z. B. auch Stolz 2007, 102 problematisiert).[5] Diese Einseitigkeit durch mündliche Paralleltexte auszugleichen ist nicht leicht. Das Ideal des konzeptionell mündlichen Textes, der spontane Dialog am Küchentisch, wird leider in der Regel nicht gedolmetscht. Ein Ansatz in dieser Richtung ist jedoch das *Court Interpreting Corpus at the University of Tampere (CoInCoUT)*, das aus finnisch-russisch gedolmetschten Gerichtsdialogen besteht (Michajlov/Isolahti 2008). Hier handelt es sich tatsächlich sowohl beim Original als auch bei der Übersetzung um weitgehend spontan gesprochene Sprache, und beide Teile sind Bestandteil des Dialogs. Jedoch ist die Äquivalenz manchmal nur dem Inhalt, nicht unbedingt dem Wortlaut nach gegeben (was man schon daran sieht, dass nicht alles, was gesprochen wird, auch übersetzt wird, vgl. ebd.). Die Gewinnung dieser Daten ist mit einem hohen Zeitaufwand verbunden, von der Erlangung einer Genehmigung aller Beteiligten zur Aufzeichnung und Verwendung der Dialoge bis zu deren Transkription, die u.a. durch sich zum Teil überlagernde Gesprächsanteile im Stil der Simultanübersetzung erschwert wird. Nicht zuletzt ist die Bandbreite an Gesprächsthemen und an Redestilen (Befragung, Zeugenaussage, Plädoyer …) relativ begrenzt.

So entstand die Idee, Film-Untertitel als Parallelkorpus zu nutzen. Diese liegen bereits graphisch vor, müssen also nicht mehr transkribiert werden, und umfassen eine ebenso große Bandbreite an Themen und Redestilen wie die ausgewählten Filme. In der Regel hat man es zwar mit fiktionalen Dialogen zu tun, die

---

Rest auf Transkripte von Fernsehinterviews, Filmdialogen u.Ä. (Grišina/Savčuk 2009, 133 f.).

5 Eine gewisse Ausnahme bildet *Europarl*, dessen Texte – zumindest in der Originalsprache – immerhin ursprünglich phonisch vorgetragen wurden, die konzeptionell jedoch in der „Sprache der Distanz" abgefasst sind. Das *Multext-East*-Korpus umfasst immerhin ein winziges Teilkorpus mit 200 englischen gesprochenen Sätzen und ihren Übersetzungen (http://nl.ijs.si/ME/V4/). An dieser Lösung ist jedoch unbefriedigend, dass die Übersetzungen von den das Korpus erarbeitenden LinguistInnen selbst oder in deren Auftrag angefertigt wurden, so dass es sich bei den so entstandenen Übersetzungen nicht um gesprochene Sprache handelt, sondern um etwas, was nach Meinung der LinguistInnen gesprochene Sprache repräsentiert, so dass linguistische Untersuchungen über gesprochene Sprache auf dieser Grundlage in einem Zirkelschluss zu enden drohen.

unvorbereitetes Sprechen lediglich imitieren, aber zumindest in Dokumentarfilmen findet man auch einige Arten ‚echter' unvorbereiteter Rede.

Diese Idee war allerdings nicht völlig neu. Schon Cysouw/Wälchli (2007: 97) erwähnen Untertitel beiläufig als "a possibly interesting source of M[assive] P[arallel] T[ext]s" – und besonders interessant sei dies, "because most of the text is direct speech". Diese Idee steht im Zusammenhang der Gewinnung „massiver" Korpora aus sehr großen Textmengen. Noch im gleichen Jahr berichtet Tiedemann (2007), dass er im Rahmen des Projekts *OPUS* bereits ein riesiges Korpus von Untertiteln zusammengestellt habe: "roughly 23,000 pairs of aligned subtitles covering about 2,700 movies in 29 languages". Diese stammen von der Datenbank *OpenSubtitles* (http://www.opensubtitles.org/), über welche die Benutzer Untertitel zu Filmen austauschen. Die Menge auch der slavischen hier verfügbaren Daten ist imposant,[6] jedoch ist ein großes Problem, dass die Untertitel von jedem hochgeladen werden können. Vermutlich sind sie zum Teil Raubkopien professionell hergestellter Untertitelübersetzungen, in den meisten Fällen aber wohl von Fans der betreffenden Filme selbst übersetzt. Das Formular zum Hochladen von Untertiteln enthält sogar ein eigenes Kästchen „Untertitel wurden maschinell übersetzt". Tiedemann (2007: 2) hat das Problem der mangelnden Qualitäts-standards zwar gesehen, dessen Lösung aber auf später ver-schoben:

> The database contains a lot of repeated subtitles; i.e. created for the same movie in the same language. We simply took the first one in the database and dismissed all the others. In future, we like to investigate possible improvements by other selection principles, e.g. taking download counts or user ratings into account.

Hier macht sich der Nachteil „massiver" Korpora bemerkbar, die zu groß sind, als dass die Texte persönlich in Augenschein genommen werden und manuell bearbeitet werden könnten. Sinclair (2001: xi) hebt hervor, dass für viele Zwecke die Methode der "early human intervention (EHI)" effizienter ist, für die kleine, aber sorgfältig zusammengestellte und auf eine bestimmte Untersuchung hin zugeschnittene Korpora am sinnvollsten sind. Wenn es also bei Untertiteln nicht um Masse geht, sondern um Klasse, führt wohl kein Weg daran vorbei, die von

---

[6] Am 21. März 2011 konnte man 151 611 polnische, 74 175 serbokroatische (40 478 serbische, 26 450 kroatische und 7 247 bosnische), 63 543 tschechische, 43 567 bulgarische, 25 533 slovenische, 10 828 russische, 8 828 slovakische, 906 makedonische und 301 ukrainische (aber offenbar keine weißrussischen oder sorbischen) sowie u.a. 148 647 englische und 9 954 deutsche Untertitelspuren herunterladen. Darunter sind z. B. 367 Untertitelspuren für den ersten *Harry-Potter*-Film, davon allein 57 auf Tschechisch.

professionellen ÜbersetzerInnen erstellten Untertitel auf gekauften DVDs selbst auszulesen. Wie dies funktioniert, sei im Folgenden kurz erläutert.

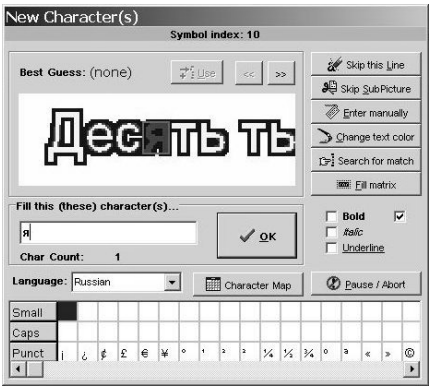
5. Untertitel ‚rippen'

Untertitel sind leider nicht als frei verfügbare Textdateien auf DVDs abgelegt, die man einfach auf seine Festplatte kopieren kann. Das erste Problem, mit dem man konfrontiert ist, ist der Kopierschutz, mit dem die meisten im Handel erhältlichen DVDs versehen sind. Seit 2003 ist es in Deutschland durch § 95a des Gesetzes über Urheberrecht und verwandte Schutzrechte (UrhG) grundsätzlich verboten, einen solchen Kopierschutz „ohne Zustimmung des Rechtsinhabers" zu umgehen.[7] Allerdings gilt die Strafandrohung von § 108b Abs. 1 nur, „wenn die Tat nicht ausschließlich zum eigenen privaten Gebrauch des Täters oder mit dem Täter persönlich verbundener Personen erfolgt". Die Zugänglichmachung von Textteilen zu Zwecken der Forschung – etwa für registrierte BenutzerInnen eines Korpus – ist aber einschließlich der dazu erforderlichen Kopien nach § 52a ausdrücklich erlaubt. Ob diese Erlaubnis zur Vervielfältigung auch die Umgehung eines etwaigen Kopierschutzes beinhaltet, bleibt jedoch unklar.

An im Internet zum Teil kostenlos verfügbaren DVD-Rippern, also Programmen, die kopiergeschützte DVDs kopieren können, herrscht kein Mangel (obwohl deren Herstellung, Verkauf usw. nach § 95a Abs. 3 UrhG rechtswidrig ist – nicht jedoch deren nichtgewerblicher Besitz). Eines dieser Programme ist die Freeware *DVDfab HD Decrypter*, mit der ich gearbeitet habe.

Hat man die gewünschten Dateien von einer nicht kopiergeschützten DVD auf die Festplatte kopiert oder mit Genehmi-gung des Urhebers den Kopierschutz umgangen, wartet schon das nächste Problem: Die Untertitel sind innerhalb der Filmdatei nicht als Text, sondern als Grafiken gespeichert (so dass der DVD-Player diese Grafiken lediglich ins Bild einblenden und nicht etwa noch Schriftarten zur Darstellung von Text bereit halten muss). Das heißt, dass man sie, ähnlich wie ein eingescanntes Textdokument, per Texterkennung *(Optical Character Recognition, OCR)* zunächst in eine Textdatei umwandeln muss. Allerdings ist der große Vorteil gegenüber der Erkennung von vom Papier eingescannten Texten, dass die

[7] Die folgenden Ausführungen gelten ebenso für Österreich und die Schweiz, wo sie in §§ 90c, 91 und 42 des österreichischen Bundesgesetzes über das Urheberrecht an Werken der Literatur und der Kunst und über verwandte Schutzrechte (UrhG) bzw. in Art. 39a, 69a und 19 des schweizerischen Bundesgesetzes über das Urheberrecht und verwandte Schutzrechte (URG) festgelegt sind.



Untertitel-Grafiken digital erstellt wurden. Daher gibt es keine Verunreinigungen oder Rasterungseffekte, die die Erkennung erschweren könnten, sondern alle Exemplare eines Zeichens haben immer exakt das gleiche Pixelmuster. Das Public-Domain-Programm *SubRip*, das genau für diesen Zweck geschrieben wurde, arbeitet daher so, dass es die BenutzerIn bei jedem noch unbekannten Pixelmuster fragt, was für ein Zeichen das sei (vgl. Screenshot). Alle anderen Exemplare dieses Pixelmusters erkennt das Programm dann selbständig.

Am Ende kann man den fertigen Text der Untertitelspur in verschiednen Formaten als Textdatei speichern. Alle diese Formate enthalten zusätzlich zum eigentlichen Untertiteltext die genaue Ein- und Ausblendzeit des Titels (und manchmal noch eine laufende Nummer). Dies erleichtert die Parallelisierung *(alignment)* der verschiedensprachigen Untertitel im Parallelkorpus erheblich.[8] Zur Illustration sei ein Ausschnitt des Outputs für die tschechischen Untertitel des Films *Casanova* im *SubRip*-eigenen Format (.srt) angeführt (links), aus dem sich durch einfache Ersetzungen sehr leicht ein korpuskompatibles Format erstellen lässt – hier etwa das in *ParaSol* benutzte (rechts):

[8] Die Synchronisierungsprobleme, die Tiedemann (2008) beschreibt, treten vor allem bei Untertiteln aus Internet-Datenbanken auf, bei denen die Benutzer individuell – und oftmals nicht sehr sorgfältig – festlegen, wann der jeweilige Untertitel ein- und ausgeblendet werden soll. Auf DVDs sind meist alle Untertitel von derselben Firma hergestellt worden, die die Ein- und Ausblendzeiten zentral festgelegt hat. Abweichungen ergeben sich hier lediglich zwischen den Übersetzungen und der Untertitelspur in der Originalsprache, da in Ersteren zusätzlich zum Ton auch im Bild zu sehende Aufschriften übersetzt werden (wobei es hier wiederum zu Unter-schieden zwischen den Übersetzungssprachen kommen kann, wenn etwa ein im Bild zu sehender englischer Eigenname in den russischen Untertiteln ins Kyrillische transkribiert wird, aber keines polnischen Untertitels bedarf).

(1)

151 00:13:32,287 --> 00:13:35,597 – Budeme se ženit. – Blahopřeji. A s kým?	<s id="150">– Budeme se ženit . – Blahopřeji . A s kým ?</s>
152 00:13:35,687 --> 00:13:38,599 Ještě nevíme. Byl tu někdo?	<s id="151">Ještě nevíme . Byl tu někdo ?</s>

Zu beachten ist dabei, dass das <s>-Tag hier im Gegensatz zu anderen Korpustexten nicht einem syntaktischen Satz ("sentence") entspricht, sondern eher als "segment" zu interpretieren ist (vgl. die Benutzung von <seg> bei Vitas/Krstev 2009, 632). Im obigen Beispiel ist bereits zu sehen, dass mehrere Repliken in einem Untertitel und damit in einem <s>-Tag zusammengefasst sein können; der folgende Ausschnitt aus den originalsprachigen Untertiteln vom Anfang des Films *Ironija sud'by* stellt ein Beispiel eines auf vier Untertitel verteilten Satzes dar:

(2)
<s id="0">Подмосковные деревни — Тропарево , Чертаново ,</s>
<s id="1">Медведково , Беляево-Богородское и , конечно же , Черемушки , —</s>
<s id="2">не подозревали о том , что обретают бессмертие</s>
<s id="3">в те грустные для них дни , когда их навсегда сметали с лица земли .</s>

Für das Tagging, d.h. die morphosyntaktische Analyse, dürfte es daher je nach angewandtem Algorithmus sinnvoll sein, die aus der Untertitelabfolge automatisch erstellten <s>-Tags zunächst zu ignorieren und nur zur Parallelisierung zu verwenden.

Leider wird man im Handel kaum DVDs finden, die Untertitel in allen größeren slavischen Sprachen enthalten. Slavische Filme werden, wenn überhaupt Untertitel vor-handen sind, meist nur mit Untertiteln in einigen nichtslavischen Sprachen ausgeliefert. Untertitel in einer ganzen Reihe slavischer Sprachen findet man i. d. R. ebenso, wie dies schon v. Waldenfels (2006: 124) für gedruckte Texte festgestellt hat, lediglich bei internationalen Bestsellern: "unfortunately, they all seem to be translated from one language: English". Dies gilt für die hollywoodzentrierte Filmindustrie in noch größerem Maße als für den Buchhandel. Jedoch findet man nur schwer DVDs, auf denen sowohl die zweit- als auch die drittgrößte slavische

Sprache – Polnisch und Ukrainisch – vertreten sind, allein schon deshalb, weil Polen den DVD-Regionalcode 2 hat, die Ukraine aber zur Region 5 gehört.[9]

6. Medienwechsel

Kommen wir nun zu der Frage, wie ‚mündlich' Untertitel sind. Die Verschriftung des gesprochenen Filmdialogs ist zwar zunächst einmal nur ein medialer und kein konzeptioneller Wechsel, jedoch sind Untertitel natürlich kein sprachwissenschaftliches Transkript. Da Untertitel in erster Linie schnell und leicht lesbar sein sollen, werden sie in mancherlei Hinsicht angepasst.
Dabei ist vor allem an Kürzungen zu denken. So werden Elemente, die eine Funktion in erster Linie für den mündlichen Dialog (im Folgenden *OT – Originalton*) haben, wie Eröffnungs- und Rückmeldungspartikeln und andere Gliederungssignale, Interjektionen oder Ausdrücke in Vokativfunktion, im Untertitel *(UT)* gern ausgelassen:

(3) *OT:* – Живо к капитану!
– Угу, понял.
UT: Живо к капитану! (*Piraty XX veka*, 00:52:26)

(4) *OT* **Harry,** what are you doing here?
UT: What are you doing here? (*Dante's Peak*, 00:05:49)

Wiederholungen werden bei der Untertitelung meist nicht wiedergegeben:

(5) *OT:* Ной! **Ной!** Где ты? **Ной!**
UT: Ной! Где ты? (*Piraty XX veka*, 00:45:07)

(6) *OT:* I have some chips and **some** crackers and **some** drinks back here.
UT: I have some chips and crackers and drinks back here. (*Dante's Peak*, 01:26:59)

[9] Russische Untertitel sind jedoch auf vielen in Mitteleuropa verkauften DVDs enthalten, da die Region 2 z.B. auch die baltischen Staaten, Georgien und nicht zuletzt Deutschland mit seinen knapp 3 Millionen russischen MuttersprachlerInnen (vgl. Brehmer 2007, 167) umfasst.

Bisweilen wird auch innerhalb des Satzes syntaktisch gekürzt, wo dies ohne Sinnverlust möglich ist:

(7) *OT:* Я тоже удивляюсь, что она выбрала тебя, **когда ты такой болван**.
UT: тоже удивляюсь, что она выбрала тебя, **такого болвана**. (*Ironija sud'by*, 00:20:46)

Seltener werden auch ganze Sätze oder Teilsätze, wenn sie für das Verständnis des Films nicht unbedingt erforderlich sind, um der Kürze willen gestrichen:

(8) *OT:* I parked my car next to a very beautiful 1964 Aston Martin. **And I'm ashamed to say** I nicked the door.
UT: I parked my car next to a very beautiful 1964 Aston Martin. I nicked the door. (*Casino Royale*, 00:29:56)

Deutlich mündliche Ausdrucksweisen werden manchmal durch schriftsprachlichere ersetzt, insbesondere wenn Letztere wie in (9) außerdem den Vorteil der Kürze oder wie in (10) in den Normvorstellungen einen zweifelhaften Status haben:

(9) *OT:* We**'re gonna** be camped here for as long as it takes.
UT: We**'ll** be camped here for as long as it takes. (*Dante's Peak*, 00:23:20)
(10) *OT:* **Жень**, **щас** серьёзно.
UT: **Женя**, **сейчас** серьезно. (*Ironija sud'by*, 00:24:04)

Typisch für schriftliche Texte ist außerdem, dass Zahlen in Ziffern wiedergegeben werden, wobei wie (11) schon einmal ein stilistischer (*dollars* vs. *bucks*) oder lexikalischer Unterschied (*a* vs. *one*) verloren gehen kann.

(11) *OT:* Now, back in **nineteen-eighty**, I would've bet **a million bucks**…
UT: Now, back in **1980**, I would've bet **$1 million**… (*Dante's Peak*, 00:22:44)

In der Häufigkeit der Abweichungen sind deutliche Unterschiede zwischen den Filmen festzustellen. Die Passage in (12), in der die fett gedruckten Textteile in der Untertitelung ersatzlos fehlen, ist für den Film *Ironija sud'by* insgesamt charakteristisch:

(12) *OT:*

Галя: Когда люди поют, **а?**	Галя: Балда.
Женя: Поют?	Женя: **Почему?**
Галя: **Гм.**	Галя: Знаешь, когда люди поют?
Женя: Когда? На демонстрациях.	Женя: Когда нет слуха и голоса?
Галя: **Так.** А еще?	Галя: **Гм-гм [= нет].**
Женя: **Ну, я не знаю.** В опере поют.	Женя: А когда?
Галя: Ну, нет…	Галя: Когда они счастливы.
Женя: Когда выпьют **поют**.	

Hingegen habe ich im gesamten James-Bond-Film *Casino Royale* nur ganze vier deutliche Abweichungen finden können. Auch gesprochene Formen wie *you got*, *yeah* und *I'm gonna* werden dort nicht zu *you've got* (oder *you have*), *yes* bzw. *I'm going to* (oder *I will*):

(13) *UT:* – You got everything you need?
– Yeah, I have.
I'm gonna get a few more photos… (*Casino Royale*, 00:53:31)

Zum Teil hat dies sicherlich mit dem Genre-Unterschied zu tun, da die geringere Dialogdichte im Actionfilm eine genauere Wiedergabe ermöglicht als die bisweilen rasend schnellen Redewechsel in einer auf Dialogwitz basierenden Liebeskomödie. Allerdings ist dieser Unterschied, anders als das Stereotyp des James-Bond-Films glauben machen könnte, eher graduell als kategorisch (7,6 Untertitel pro Minute in *Casino Royale* gegenüber 10,6 in *Ironija sud'by*). Zudem gilt die insgesamt geringere Abweichung von der Tonspur auch für andere im Original englischsprachige Filme, während der sowjetische Actionfilm *Piraty XX veka* (der sogar mit nur 4,0 Untertiteln pro Minute auskommt) relativ viele Abweichungen zwischen Ton und Titel enthält. Daher ist die Erklärung wohl vor allem darin zu suchen, dass manche Untertitel gar nicht in erster Linie zur Einblendung im Film gedacht sind. Die primäre Funktion der originalsprachigen Untertitel auf DVDs mit professionell (um nicht zu sagen: industriell) hergestellter mehrsprachiger Untertitelung ist vielmehr, als Übersetzungsvorlage für die fremdsprachigen Untertitel zu dienen.[10] Daher geht Vollständigkeit, die den ÜbersetzerInnen hilft, den Kontext zu erfassen und eine möglichst adäquate Übersetzung anzufertigen, letztendlich wohl

[10] Ein Vergleich der auf einigen im Handel erhältlichen DVDs enthaltenen englischen Untertitel mit der Übersetzungsvorlage, die der Übersetzerin von der Firma Deluxe Digital Studios Subtitling and Localization zur Anfertigung der deutschen Untertitel zur Verfügung gestellt wurde, ergab deren völlige Identität.

doch vor Kürze und schneller Erfassbarkeit. Um die schnelle Lesbarkeit müssen sich dann theoretisch die ÜbersetzerInnen kümmern, die aber ebenfalls, wie wir im Folgenden sehen werden, insgesamt erstaunlich genau übersetzen.

7. Übersetzung

Für die Herstellung fremdsprachiger Untertitel kommen theoretisch mehrere Quellen in Betracht. So befinden sich die Dialoge ja bereits im Drehbuch, das aber häufig von der tatsächlichen Umsetzung am Set abweicht. Viele der bei *OpenSubtitles* erhältlichen, von Laien angefertigten Untertitel sind direkt von der Tonspur abgehört und übersetzt. Im professionellen Bereich erweist es sich aber klar als effizienter und weniger fehlerträchtig, wenn in einem ersten Schritt eine MuttersprachlerIn der Originalsprache die zu hörenden Dialoge abschreibt und diese Schriftfassung dann in einem zweiten Schritt von einer MuttersprachlerIn der Zielsprache übersetzt wird. Dies gilt insbesondere, wenn von vornherein Übersetzungen in mehrere Sprachen geplant sind, da die Arbeit des Abhörens der Tonspur ja sonst mehrfach anfallen würde. Diese Vorgehensweise schließt nicht aus, dass die ÜbersetzerIn beim Übersetzen der schriftlichen Untertitel zusätzlich die Tonspur in Betracht zieht. So hat etwa die serbische ÜbersetzerIn den in (8) **Error! Reference source not found.**ausgelassenen Teilsatz (der auch in den kroatischen, slovenischen und polnischen Untertiteln fehlt) in ihre Übersetzung eingefügt:

(14) *e.* I parked my car next to a very beautiful 1964 Aston Martin. I nicked the door
sb. …i parkirao sam kola pored jednog divnog aston martina iz 1964. **Sramota me je, ali** ogrebao sam vrata. (*Casino Royale*, 00:29:56)

Ansonsten ergibt eine genauere Betrachtung der Untertitel-Übersetzungen eigentlich keine wesentlichen Unterschiede zu anderen literarischen Übersetzungen. Natürlich gibt es auch hier Übersetzungsfehler. Zwei besonders deutliche Beispiele aus den russischen Untertiteln zum Film *Casanova* seien hier aufgeführt. In (15) führt ein Säufer ein schlagendes Argument an, um zu belegen, dass ihn niemand bestochen habe, – welches die russische ÜbersetzerIn aber leider nicht verstanden hat:

(15) *e.* Nobody bribed me. Would I be sober if they had?
r. Никто меня не подкупал. Я бы устоял перед искушением? (*Casanova*, 00:37:15)

In (16) hat die russische ÜbersetzerIn *say* mit *tell* verwechselt und außerdem *madman* als *madam* gelesen:

(16) *e.* – […] Bernardo Guardi, who was said to live in this very city.
– Bernardo Guardi. That madman.
r. – […] Бернардо Гуарди, которому повелели тут жить.
– Бернардо Гуарди. Этой мадам. (*Casanova*, 00:09:02)

Auch wenn diese extremen Aussetzer kaum zu entschuldigen sind, ist doch anzumerken, dass die Filmfirmen, die die Untertitelung in Auftrag geben, es den ÜbersetzerInnen nicht gerade leicht machen. Zum Zeitdruck kommt noch erschwerend hinzu, dass der audiovisuelle Kontext der Untertitel, den man bisweilen zur Übersetzung benötigt, oft nur schwer zu erkennen ist, weil den ÜbersetzerInnen aus Angst, dass der Film noch vor seinem Verkaufsstart illegal in Umlauf kommen könnte, nur ein verzerrtes Bild und ein Ton schlechter Qualität zur Verfügung gestellt wird. Vor diesem Hintergrund ist erstaunlich, welche Leistungen die ÜbersetzerInnen zum Großteil erbringen. Ein Beispiel dafür, aus demselben Film, bietet die kroatische Übersetzung: In (17) stellt sich ein junger Mann beim Versuch, Casanova wegen dessen Avancen gegenüber seiner Angebeteten zum Duell zu fordern, so ungeschickt an, dass dieser versehentlich auf den ihm hingeworfenen Fehdehandschuh tritt. Das Wortspiel ist im Kroatischen geradezu kongenial wiedergegeben:

(17) *e.* You have sullied my glove. I mean, love. My love.
kr. Ukaljali ste mi rukavicu! Ljubavnicu. Moju ljubav. (*Casanova*, 00:19:10)

Insgesamt kann man also bei Untertitel-Übersetzungen in ähnlicher Weise von weitestgehend gegebener Äquivalenz (bzw. Adäquatheit) ausgehen wie bei anderen literarischen Übersetzungen auch.

8. Fiktionalisierung

Was unterscheidet nun ein aus Untertiteln gewonnenes Korpus von einem ‚richtigen' Korpus gesprochener Sprache? Auf den ersten Blick relativ wenig. Vergleicht man etwa die Suchergebnisse im mündlichen Teil des Russischen Nationalkorpus (RNK) mit denen in einem winzigen Testkorpus, das im Rahmen

der *ParaSol*-Architektur aus den Untertiteln zu den Filmen *Ironija sud'by* (14 298 Tokens im Russischen) und *Casanova* (9 181 Tokens) erstellt wurde, so ist der auffälligste Unterschied die Zeichensetzung: Während die von LinguistInnen angefertigten Transkripte kurze und lange Pausen durch „/" bzw. „//" kennzeichnen, enthalten die Untertitel orthographische, syntaktisch gesetzte Kommata. Außerdem fehlt im Untertitelkorpus der in Transkripten häufige Hinweis „[нрзб]" (für *неразборчиво* 'unverständlich'). In manchen mündlichen Korpora werden noch weitere phonetische Begebenheiten in der Transkription berücksichtigt (vgl. z.B. Kibrik/Podlesskaja 2003). Vor allem werden bei der Transkription die Sprecher identifiziert. Da man beim Lesen von Untertiteln aber normalerweise im Bild sieht, wer gerade spricht,[11] fehlen diese Sprecher-identifikationen hier – was das Verfolgen eines Dialogs zu einem unbekannten Film in einem Untertitelkorpus manchmal nicht gerade einfacher macht. In transkribierten Korpora werden auch bisweilen Teile des außersprachlichen Kontextes erläutert, während man bei Untertiteln dazu auf die audiovisuellen Informationen angewiesen ist.[12]

Andererseits enthalten auch klassische ‚mündliche' Korpora in der Regel keine Lautschrift und auch wenig Ansätze zur Darstellung der realen Aussprache mit Mitteln der Orthographie, da auch sie einigermaßen lesbar bleiben sollen. So findet man im mündlichen RNK z. B. 778-mal „50" in Ziffern geschrieben[13], 1 192-mal orthographisch korrektes „пятьдесят" und nur je ein einziges Mal „псят" und „пьсят", die der wohl häufigsten Aussprache (jedenfalls in nicht-finaler Position) am nächsten kommen, sowie 12-mal „писят" und 6-mal „пидисят". Aus Romanen als *eye dialect* bekannte Schreibweisen, die daher im Gegensatz zu den Verkürzungen von '50'[14] auch im schriftlichen Basis-Korpus des RNK vorkommen, haben es etwas leichter: So wird z.B. 7 473-mal „щас" transkribiert (sowie 62-mal „счас" und 24-mal „сичас"), aber 22 353-mal bleibt es bei „сейчас", 8 931 „чё" stehen 10 756 „чего" gegenüber, 1 345 „прям" 2 248 „прямо", 979 Formen von

---

11 Text nicht im Bild sichtbarer Personen – etwa vom anderen Ende der Telefonleitung oder bei einem Erzähler aus dem Off – wird in Untertiteln meist durch Kursivschrift kenntlich gemacht.

12 Dieser Umstand spricht für kleine Untertitel-Korpora, bei denen es realistisch ist, im Zweifelsfall den Film zu Rate zu ziehen. Bei der riesigen Untertitel-Datenbank im Rahmen des *OPUS*-Projekts ist das kaum möglich.

13 Die Ziffernschreibweise von *pjat'desjat* ist noch deutlich häufiger, denn auch die 69 Vorkommen von „55", die 111 Mal „150", die 5 Mal „1954" usw., die alle als separate Wörter behandelt werden, müssten ja mitgezählt werden.

14 Im Grundkorpus des RNK ist *p'sjat* ein einziges Mal zu finden (in einer Erzählung von Vjačeslav Rybakov). Die anderen Formen liefern je null Treffer (außer *psjat*, das als Genitiv Plural von *psënok* 'Welpe; Köterchen' vorkommt).

„тыщ*“ („тыща“, „тыщи“, „тыщ“ usw.) 2 919 „тысяч*“, und sogar die von der realen Aussprache sehr weit entfernte Schreibung „здравствуй(те)“ hat im mündlichen Korpus mit 4 810 Vorkommen ein klares Übergewicht gegenüber den 2 190 Varianten von „здрас*“ („здрасьте“, „здрасти“ usw.). In den auf schnelles Lesen angelegten Untertiteln kommt solcher *eye dialect* allerdings so gut wie gar nicht vor.

In medialer Hinsicht liegen also sowohl bei Untertiteln als auch bei ‚mündlichen' Korpora *graphische* Texte vor, und der Unterschied, dass die traditionell transkribierten Texte die am Korpus arbeitenden LinguistInnen noch in phonischer Form erreicht haben, macht sich nur wenig bemerkbar, da die Texte dann doch in eine weitgehend orthographische Form gebracht werden.

Wie aber steht es um die *konzeptionelle* Mündlichkeit? Koch/Oesterreicher (1985, 23) zählen in einer offenen Liste 18 Merkmale der „Sprache der Nähe“ auf, an denen man sich zur Einschätzung der Korpustexte orientieren kann (gegliedert in „Kommunikationsbedingungen“ und „Versprachlichungsstrategien“), von denen aus Platzgründen hier nur die wichtigsten behandelt werden können. Ganz oben steht „Dialog“. Da sowohl Erzählerkommentare als auch Monologe im Stil Hamlets in modernen Filmen selten sind, kann man ein Untertitelkorpus grundsätzlich als dialogisch ansehen. Am Testkorpus lässt sich aber demonstrieren, dass diese Dialogizität sich auch sprachlich auswirkt und das Untertitelkorpus – am Beispiel des RNK – deutlich näher an mündliche als an schriftliche Korpora rückt. Zum einen sieht man das an der Häufigkeit von Fragesätzen, die ja einen wichtigen Bestandteil von Dialogen bilden, untersucht anhand der Relation von Fragezeichen und Punkten:[15]

(18)

	RNK Basis	**russ. Untertitel**	**RNK mündl.**
Anzahl Fragezeichen	1 095 436	706	271 434
Anzahl Punkte	11 839 385	2 539	878 861
Verhältnis „?“ : „.“	**1 : 10,8**	**1 : 3,6**	**1 : 3,2**
Sätze insgesamt	14 711 345	ca. 4 000	1 520 959
Frequenz Fragesätze	**7,4 %**	**ca. 17 %**	**17,8 %**

[15] Das Verhältnis dieser Satzzeichen entspricht natürlich nur näherungsweise dem Verhältnis direkter Fragen zu Aussagesätzen, da Punkte z.B. auch nach Abkürzungen vorkommen, manche Sätze völlig ohne Satzzeichen enden usw.

Auch die Frequenz eines anderen direktiven Sprechakttyps, der Bitte, lässt sich anhand der Häufigkeit des Wortes *požalujsta* ‘bitte’ leicht abschätzen. Hier liegt das Untertitelkorpus ebenfalls erheblich näher am mündlichen als am schriftlichen Vergleichskorpus.

(19)

	RNK Basis	**russ. Untertitel**	**RNK mündl.**
Tokens insgesamt	176 226 551	23 479	9 526 425
Anzahl *požalujsta*	15 232	13	10 210
Frequenz	$1/11569$	$1/1806$	$1/933$

Einige von Koch/Oesterreichers Kriterien lassen sich nur schwer abschätzen. „*Face-to-face*-Interaktion“ ist in Filmen, abgesehen von Telefongesprächen oder der Verfilmung des Schreibens und Lesens eines Briefs, wohl meistens gegeben, während schriftliche Korpora außerhalb der Wiedergabe direkter Rede durch raumzeitliche Trennung des Autors oder Erzählers vom Leser gekennzeichnet sind. Über den Grad der „Vertrautheit der Partner“ geben die Korpusdaten nur wenig Aufschluss.[16] „Spontaneität“ und „freie Themenentwicklung“ werden in fiktionalen Filmen „mimetisch-imitativ“ (vgl. Koch/Oesterreicher 1985, 24) gespielt.[17]
Wenn man die Häufigkeit des Präteritums gegenüber dem Präsens als Anzeichen für die „Situationsentbindung“ werten will, zeigt sich (anhand des relativ häufigen Modalverbs *xotet’* ‘wollen’), dass das Untertitel-Testkorpus sogar eine größere referenzielle Nähe aufweist als das mündliche RNK, was dadurch zu erklären ist, dass in das RNK auch distanzsprachlichere Textsorten wie Vorlesungen,

[16] Da die distanzierte pronominale Anrede im Russischen mit der vertrauten Anrede der 2. Person Plural zusammenfällt, sind Vergleiche von *ty* vs. *vy* u.Ä. wenig aussagekräftig. Für die nominale Anrede, für die man im RNK feststellen kann, dass z.B. *Sergej* im schriftlichen Korpus 4,9-mal häufiger ist als *Serëža*, im mündlichen aber nur 2,5-mal häufiger, ist das bisher erstellte Testkorpus noch viel zu klein, um aussagekräftige Ergebnisse auch bei den Untertiteln zu erhalten.

[17] Dies gilt auch für einen Teil des mündlichen RNK, das laut Savčuk (2009, 589 f.) zu 12 % aus transkribierten Kinofilmen besteht, die fast alle fiktional sind.
Im Gegensatz zu den in Fn. 5 erwähnten Übersetzungen englischer Gespräche im *Multext-East*-Korpus handelt es sich bei dieser imitierten Mündlichkeit jedoch nicht um die Vorstellungen von LinguistInnen über spontan gesprochene Sprache, sondern um die von DrehbuchschreiberInnen, RegisseurInnen und SchauspielerInnen sowie im Fall der übersetzten Untertitel von ÜbersetzerInnen, die für die Filmindustrie arbeiten. Dadurch kann die Linguistik immerhin ein außerhalb ihrer selbst gelegenes Phänomen untersuchen und unterliegt nicht dem Zirkelschluss, ihre eigenen Vorstellungen zu beschreiben.

(Nach-)Erzählungen, öffentliche Diskussionen oder Interviews eingegangen sind (Savčuk 2009, 590).

(20)

	RNK Basis	**russ. Untertitel**	**RNK mündl.**
Präsens von *xotet'*	103 258	42	16 362
Präteritum von *xotet'*	92 916	8	7 957
Verhältnis Präs. : Prät.	**1,1 : 1**	**5,3 : 1**	**2,1 : 1**

Zu den „Versprachlichungsstrategien" (Koch/Oesterreicher 1985, 23) haben wir bereits gesehen, dass beim Medienwechsel von der Tonspur zum Untertitel die Informationsdichte und Kompaktheit erhöht werden und dadurch sicherlich an das Niveau geschriebener Texte heranreichen. Die Häufigkeit der Wörter *nu*, *a* und *da*, die in gesprochener Sprache oft als Eröffnungspartikeln fungieren (Rathmayr 1985, 166–174), in geschriebener aber fast ausschließlich in anderen Funktionen und dementsprechend seltener verwendet werden, ist in meinem kleinen Testkorpus sogar noch geringer als im Basis-Korpus des RNK. Dies ist wohl darauf zurückzuführen, dass an die Kompaktheit von Untertiteln hohe Ansprüche gestellt werden (i.d.R. max. 2 Zeilen, bei schnellem Gespräch möglichst noch kürzer), während in Romanen genug Raum für die Stilisierung von Mündlichkeit in direkter Rede ist. Außerdem kommen phatische Partikeln, Füllwörter u.Ä. sicherlich schon auf der Tonspur seltener vor als in realen Gesprächen, da der Filmtext ja in Wirklichkeit auswendig gelernt ist.

(21)

	RNK Basis	**russ. Untertitel**	**RNK mündl.**
Tokens insgesamt	176 226 551	23 479	9 526 425
nu	163 678 $= 1/1077$	5 $= 1/4696$	100 323 $= 1/95$
a	1 400 685 $= 1/126$	44 $= 1/534$	162 271 $= 1/59$
da	303 247 $= 1/581$	9 $= 1/2609$	87 232 $= 1/109$

Der Unterschied an „Komplexität" lässt sich an der mittleren Satzlänge der Korpora ablesen, die sich aus der jeweiligen Gesamtzahl der Tokens und der Sätze, die bereits in (18) und (19) angegeben wurden, errechnen lässt. So ergibt sich für das

Basis-RNK eine mittlere Satzlänge von 12,0 Tokens[18], für das mündliche Teilkorpus nur 6,3 Tokens. Das Untertitel-Testkorpus unterschreitet diesen Wert mit ca. 5,9 sogar noch ein wenig.

9. Fazit

Korpora von Untertiteln sind in medialer Hinsicht selbstverständlich graphische und nicht etwa phonische Texte. Konzeptionell jedoch sind sie eindeutig zur von Koch/Oesterreicher (1985) beschriebenen „Sprache der Nähe" zu zählen, auch wenn sie wegen des Primats der schnellen Lesbarkeit einzelne distanzsprachliche Elemente aufweisen. Zur Ausbalancierung eines üblicherweise weitgehend distanzsprachlichen Parallelkorpus sind sie daher gut geeignet, und auch sonst stellen sie im Rahmen der „Berücksichtigung und Erforschung der konzeptionellen sprachlich-diskurstraditionellen Abstufungen" (Koch/Oesterreicher 2007, 368) ein interessantes Beispiel gesprochen-graphischer Texte dar.

Quellen

Hallström, L. (Regie). 2005. *Casanova.* Touchstone Pictures. DVD von Buena Vista Home Entertainment, vertrieben von Continental film, Zagreb (o. J.).

Campbell, M. (Regie). 2007. *Casino Royale.* EON Productions 2006. DVD von Sony Pictures Home Entertainment, vertrieben von BLITZ film i video distribucija, Zagreb.

Donaldson, R (Regie). 2006 [1997]. *Dante's Peak / Danteov vrh.* Universal Pictures. DVD von Universal Pictures, vertrieben von Continental film, Zagreb

Rjazanov, Ė (Regie). 2001 [1975]. *Ironija sud'by ili „S lëgkim parom".* Mosfil'm DVD von Ruscico, Moskva.

Durov, B. (Regie). 2001 [1979]. *Piraty XX veka.* Kinostudija im. Gor'kogo. DVD von Ruscico, Moskva.

Sekundärliteratur

Brehmer, B. 2007. Sprechen Sie Qwelja? Formen und Folgen russisch-deutscher Zweisprachigkeit in Deutschland. In: Anstatt, T. (Hrsg.), *Mehrsprachigkeit bei Kindern und Erwachsenen: Erwerb, Formen, Förderung.* Tübingen: Attempto, S. 163–185.

[18] Im Untertitel-Testkorpus sind, ebenso wie in *ParaSol*-Korpus, auch Satzzeichen (incl. Kommata) Tokens. Im RNK sind die Gesamtzahlen mit dem Wort *slovo* angegeben, wobei nicht ganz klar ist, ob Satzzeichen hier tatsächlich ausgeschlossen sind.

Cysouw, M./Wälchli, B. 2007. Parallel Texts: Using Translational Equivalents in Linguistic Typology. In: *Sprachtypologie und Universalienforschung* 60(2): S. 95–99.

Dobrovol'skij, D./Kretov, A. /Šarov, S. 2005. Korpus parallel'nych tekstov: Architektura i vozmožnosti ispol'zovanija. In: *Nacional'nyj korpus russkogo jazyka: 2003–2005.* Moskva: Indrik, S. 263–296.

Grišina, E./Savčuk, S. 2008. Korpus zvučaščej russkoj reči v sostave Nacional'nogo korpusa russkogo jazyka. Proekt. In: Kibrik, A. et al. (Hrsg.), *Komp'juternaja lingvistika i intellektual'nye technologii. Po materialam ežegodnoj meždunarodnoj konferencii „Dialog" (Bekasovo, 4–8 ijunja 2008 g.)* 7 (14). Moskva: RGGU, S. 125–132.

Grišina, E./Savčuk, S. 2009. Korpus ustnych tekstov v NKRJa: sostav i struktura. In: Plungjan, V. (Hrsg.), *Nacional'nyj korpus russkogo jazyka: 2006–2008. Novye rezul'taty i perspektivy.*. Sankt-Peterburg: Nestor-Istorija, S. 129–149.

Kabatek, J. 2000. L'oral et l'écrit – quelques aspects théoriques d'un «nouveau» paradigme dans le canon de la linguistique romane. In: W. Dahmen et al. (Hrsg.), *Kanonbildung in der Romanistik und in den Nachbardisziplinen: Romanistisches Kolloquium XIV.* Tübingen: Narr, S. 305–320.

Kibrik, A./Podlesskaja, V. 2003. K sozdaniju korpusov ustnoj russkoj reči: Principy transkribirovanija. In: *Naučno-techničeskaja informacija, Serija 2: Informacionnye processy i sistemy* 10: S. 5–12.

Koch, P. 1999. „Gesprochen/geschrieben" – eine eigene Varietätendimension? In: Greiner, N. et al. (Hrsg.), *Texte und Kontext in Sprachen und Kulturen: Festschrift für Jörn Albrecht.* Trier: Wissenschaftlicher Verlag Trier, S. 141–168.

Koch, P./Oesterreicher, W. 1985. Sprache der Nähe – Sprache der Distanz: Mündlichkeit und Schriftlichkeit im Spannungsfeld von Sprachtheorie und Sprachgeschichte. In: *Romanistisches Jahrbuch* 36: S. 15–43.

Koch, P./Oesterreicher, W. 2007. Schriftlichkeit und kommunikative Distanz. *Zeitschrift für germanistische Linguistik* 35: S. 346–375.

Michajlov, M./Isolahti, N. 2008. Korpus ustnych perevodov kak novyj tip korpusa tekstov. In: Kibrik, A. et al. (Hrsg.), *Komp'juternaja lingvistika i intellektual'nye technologii: Po materialam ežegodnoj Meždunarodnoj konferencii „Dialog"* (14). Moskva: RGGU.

Orechov, B. 2009. Parallel'nyj korpus perevodov „Slova o polku Igoreve": Itogi i perspektivy. In: Plungjan, V. (Hrsg.), *Nacional'nyj korpus russkogo jazyka: 2006–2008. Novye rezul'taty i perspektivy.* Sankt-Peterburg: Nestor-Istorija, S. 462–473.

Rathmayr, R. 1985. *Die russischen Partikeln als Pragmalexeme.* München: Otto Sagner.

Savčuk, S. 2009. Rossijskie razrabotki korpusov ustnoj reči. In: Tošović, B. (Hrsg.), *Die Unterschiede zwischen dem Bosnischen/Bosniakischen, Kroatischen und Serbischen: Lexik – Wortbildung – Phraseologie.* Wien/Münster: LIT, S. 579–595.

Simeon, I. 2002. Paralelni korpusi i višejezični rječnici. *Filologija* 38/39, S. 209–215.

Sinclair, J. 2001. Preface. In: Ghadessy, M. (Hrsg.), *Small Corpus Studies and ELT: Theory and Practice.* Amsterdam: John Benjamins, S. vii–xv.

Stolz, T. 2007. *Harry Potter* meets *Le petit prince* – On the Usefulness of Parallel Corpora in Crosslinguistic Investigations. In: *Studien in Typologie und Universalienforschung* 60(2): S. 100–117.

Tadić, M. 2000. Building the Croatian-English Parallel Corpus. In: Gavrilidou, M. et al. (Hrsg.), *Proceedings of the 2nd International Conference on Language Resources and Evaluation, Athens, Greece, 31 May – 2 June 2000 (LREC 2000).* Bd. 1. Paris: ELRA, S. 523–530.

Tiedemann, J. 2007. Building a Multilingual Parallel Subtitle Corpus. In: Dix, P. et al. (Hrsg.), *Proceedings of the 17th Meeting of Computational Linguistics in the Netherlands (CLIN 17)*. Utrecht: Lot.

Tiedemann, J. 2008. Synchronizing Translated Movie Subtitles. In: *Proceedings of the 6th International Conference on Language Resources and Evaluation, Marrakech, Morocco, 26–31 May 2008 (LREC 2008)*. Paris: ELRA.

Vitas, D./Krstev, C. 2009. O paralelnim korpusima, a posebno o beogradskim paralelnim korpusima i načinu njihove eksploatacije. In: Branko Tošović (Hrsg.), *Die Unterschiede zwischen dem Bosnischen/Bosniakischen, Kroatischen und Serbischen: Lexik – Wortbildung – Phraseologie.* Wien/Münster: LIT, S. 630–649.

von Waldenfels, R. 2006. Compiling a Parallel Corpus of Slavic Languages: Text Strategies, Tools and the Question of Lemmatization in Alignment. In: Brehmer, B. et al. (Hrsg.), *Beiträge der Europäischen Slavistischen Linguistik (Polyslav) 9*. München: Otto Sagner, S. 123–138.

Wälchli, B. 2007. Advantages and Disadvantages of Using Parallel Texts in Typological Investigations. *Studien in Typologie und Universalienforschung* 60(2): S. 118–134.

Abstract

In contrast to monolingual corpora, for which transcripts of oral communication are nowadays increasingly available, parallel corpora are still seriously biased towards written texts. An approach to counter this bias might be the inclusion of movie subtitles, which can be extracted from DVDs. Therefore this paper examines in how far subtitles represent orality, which is understood in the sense of "language of proximity" in Koch/Oesterreicher's (1985) framework. For the analysis the creation of multilingual subtitles is broken down into three steps: the fictionalization of movie dialogue in contrast to real dialogue; the change of medium from the sound track to the subtitle; and the translation from the original language to the target languages. The result of this analysis on the basis of a tiny test corpus of subtitles is that despite some adaptations for better readability, subtitles exhibit the "language of proximity" to a great extent and can therefore provide a suitable counter-balance for written parallel corpora.

Linguistische Beiträge zur Slavistik

XIX. JungslavistInnen-Treffen in Berlin

16. - 18. Dezember 2010

Herausgegeben von

Luka Szucsich, Natalia Gagarina, Elena Gorishneva, Joanna Leszkowicz

Verlag Otto Sagner · München – Berlin – Washington, D.C.

2012