SubRosa – Multi-Feature-Ähnlichkeitsvergleiche von Untertiteln

Luhmann, Jan
jan.luhmann@gmx.net
Computational Humanities Group, Universität Leipzig

Burghardt, Manuel
burghardt@informatik.uni-leipzig.de
Computational Humanities Group, Universität Leipzig

Tiepmar, Jochen
jtiepmar@informatik.uni-leipzig.de
Computational Humanities Group, Universität Leipzig

Einleitung: Filmanalyse auf Basis von Untertiteln

Mit der stetig wachsenden Verfügbarkeit von Filmen und Serien, die durch Streaming-Dienste wie Netflix und Amazon Prime in den letzten Jahren weiter befördert wurde, ergeben sich aus Perspektive der Filmanalyse ganz neue Möglichkeiten für quantitative Untersuchungen im Sinne des *distant viewing* (Arnold & Tilton, 2019). Wenngleich Film zunächst vor allem ein visuelles Medium ist, so werden in zunehmendem Maße auch Metadaten und insbesondere die Dialoge (vgl. Kozloff, 2000) in Form von online verfügbaren Untertiteln, Drehbüchern und Fan-Transkripten Gegenstand quantitativer Untersuchungen (vgl. Bednarek, 2020; Burghardt et al., 2016, 2019; Schmidt, 2014). Insbesondere die freie Datenbank *OpenSubtitles*[1] hat sich hier als ertragreiche Datenquelle bewährt. Während die Daten von *OpenSubtitles* bislang vor allem im Bereich maschineller Übersetzung (vgl. Müller & Volk, 2013; Lison & Tiedemann, 2016; Tiedemann, 2016) Verwendung fanden, schlagen wir in diesem Artikel eine Nutzung im Sinne quantitativer Filmstilanalyse basierend auf Ähnlichkeitsvergleichen vor. Wir erweitern damit bestehende Arbeiten (Blackstock & Spitz, 2008; Nessel & Cimpa, 2011; Bougiatiotis & Giannakopolous, 2017), die sich ebenfalls mit Ähnlichkeitsvergleichen von Untertiteln beschäftigen, dabei aber jeweils mit relativ überschaubaren Korpora arbeiten oder sehr spezifische Ansätze der Ähnlichkeitsberechnung umsetzen.

Wir präsentieren das experimentelle Analysetool *SubRosa*, welches Ähnlichkeitsvergleiche für mehrere tausend Untertitel über eine grafische Benutzeroberfläche erlaubt. Wir setzen dabei eine ganze Reihe von Features für die Ähnlichkeitsberechnung zwischen Untertiteln um, die zudem jeweils individuell gewichtet werden können. *SubRosa* versteht sich damit als exploratives Werkzeug, um die grundlegende Eignung unterschiedlicher Features bzw. Feature-Kombinationen für die computergestützte Ähnlichkeitsberechnung zwischen Untertiteln zu untersuchen, welche dann wiederum in einem nächsten Schritt für großangelegte Ähnlichkeitsvergleiche mithilfe statistischer Verfahren genutzt werden können.

Korpus und Datenaufbereitung

SubRosa stellt Vergleiche zwischen insgesamt 5.896 englischen Untertiteln an, die über *OpenSubtitles* bezogen wurden. *OpenSubtitles* versteht sich als offene Plattform, bei der Nutzer*innen Untertitel in unterschiedlichen Sprachen für unterschiedliche Filme hochladen können. Das Format der Untertitel entspricht dem Exportformat des *SubRip*-Tools, welches automatisiert über OCR Textzeilen aus Filmen mit bereits bestehenden Untertiteln extrahiert. Darüber hinaus werden aber auch viele von Nutzer*innen selbst transkribierte Untertitel hochgeladen. Im Ergebnis gibt es so für die meisten Filme mehrere Versionen von Untertiteln. Wir wählen jeweils die Version für unser Korpus aus, die einer automatischen Validierung in Hinblick auf Encoding- oder OCR-Fehler Stand hält. Weiterhin werden alle ausgewählten Untertitel grundlegend aufbereitet, d.h. es werden bspw. Metainformationen, Autoren-Tags, etc. entfernt, die definitiv nicht Teil des eigentlichen Filmdialogs sind. Als nächstes erfolgt eine Vorverarbeitung der Untertitel im Sinne des *natural language processing* (NLP), welche die folgenden Einzelschritte enthält: Tokenisierung, Satzsegmentierung, Lemmatisierung, POS-Tagging und *named entity recognition*. Zuletzt werden alle Untertitel mit Metadaten wie etwa „Titel“, „Jahr der Veröffentlichung“, „Genre“, etc. verknüpft, die über die IMDb-Datenbank[2] bezogen werden.

Analyseverfahren

Mit *SubRosa* setzen wir einen parametrisierbaren Ähnlichkeitsvergleich zwischen Filmuntertiteln um, der auf ganz unterschiedlichen Features basiert. Die nachfolgenden Features sind allesamt über eine interaktiven Web-Applikation verfügbar, die eine Ähnlichkeitssuche für die eingangs erwähnten annähernd 6.000 englischsprachigen Film-Untertitel erlaubt.

- **SubRosa Code**: http://github.com/bbrause/subrosa
- **SubRosa Live-Demo**: http://ch01.informatik.uni-leipzig.de:5001/

Features auf der inhaltlichen Ebene

a) Bag of words / tf-idf („Worüber sprechen die Figuren?“): Das *bag of words*-Modell ist ein einfacher Ansatz für die Repräsentation von Textdokumenten im NLP und Information Retrieval. In unserem Anwendungskontext entspricht ein Untertitel einem „Dokument“, für das die einzelnen lemmatisierten Tokens jeweils mit einer sublinearen tf-idf-Skalierung (Manning et al., 2008, S. 126-127) gewichtet werden. Durch diese Gewichtung können wir diejenigen Wörter identifizieren, die in einem bestimmten Dokument häufig vorkommen, aber insgesamt im Gesamtkorpus nur selten auftreten. Es kann davon ausgegangen werden, dass diese Begriffe für das jeweilige Dokument dann besonders aussagekräftig sind. Dementsprechend filtern wir alle Begriffe heraus, die in weniger als 2,5% und mehr als 95% aller Dokumente vorkommen. Darüber hinaus werden *named entities*, die Personen-, Orts- oder Institutionsnamen bezeichnen, entfernt, da diese die Ergebnisse stark verzerren können. Es verbleiben insgesamt 4.952 Wörter, die beim Ähnlichkeitsvergleich der Untertitel berücksichtigt werden.

b) Sentiment Analyse („Was fühlen die Figuren?“): Um Muster bzgl. der von den Figuren im Dialog zum Ausdruck gebrachten Gefühle und Emotionen automatisch zu detektieren, wurde das weitverbreitete *open source*-Tool *VADER Sentiment* (Hutto & Gilbert, 2014) verwendet. Dabei werden für beliebige Textabschnitte Sentiment-Bewertungen im Bereich -1 (maximal negativ) bis +1 (maximal positiv) berechnet. Da sich Emotionen im Laufe eines Films meist sehr divers entwickeln, ist es nicht sinnvoll, das Sentiment des gesamten Filmdialogs mit einem einzigen Wert wiederzugeben. Stattdessen berechnen wir für jede Sekunde eines Films einen spezifischen Sentiment-Wert für den dort gesprochenen Dialog, sodass sich für jeden Film eine Zeitreihe von Sentiment-Werten ergibt. Als Features dieser Zeitreihen extrahieren wir den Mittelwert und Quartilswerte, um die Verteilung der Sentiment-Werte zu erfassen. Weiterhin wird die Nulldurchgangsrate der Zeitreihenkurve sowie deren erste und zweite Ableitung ausgewertet, um Hinweise auf periodische Eigenschaften zu erlangen.

Features auf der stilistischen Ebene („Wie sprechen die Figuren?“)

a) Stoppwort-Verteilung: Als weitere Features implementieren wir eine Analyse der Verteilung von Stoppwörtern, also von Wörtern, die in unserem Korpus am häufigsten auftreten und im Gegensatz zum vorherigen Ansatz nur geringe inhaltliche Aussagekraft für einen Film besitzen. Wir berücksichtigen insgesamt 87 Stoppwörter, die nach ihrer Termfrequenz gewichtet werden.

b) POS-Trigramme: Darüber hinaus setzen wir einen Ansatz von Argamon et al. (2003) und Santini (2004) um, die im Kontext stilometrischer Genreklassifikation mit POS-Trigrammen arbeiten. Wir ignorieren dabei all die POS-Trigramme, die in weniger als 90% unserer Dokumente vorkommen, was zu insgesamt 417 verbleibenden POS-Trigrammen führt. Gewichtet werden diese ebenfalls nach ihrer Termfrequenz.

c) Statistische Maße: Wir berechnen außerdem verschiedene statistische Maße, die im Bereich der Stilometrie weit verbreitet sind und die als weitere Features bei unserer Ähnlichkeitsberechnung verwendet werden können. Zu diesen Maßen zählen die Durchschnittswerte einfacher Wort- und Satzlängen sowie auch die *Entropie* (Shannon, 1948) und die *standardized type-token ratio* (Johnson, 1944; Torruella & Capsada, 2013).

d) Dialogtempo („Wie schnell bzw. wie viel wird gesprochen?“): Als letztes Feature betrachten wir das „Dialogtempo“, das sich allerdings nicht auf die Sprechgeschwindigkeit einzelner Figuren bezieht, sondern vielmehr Dialoganteile pro Zeit misst. Analog zum Verfahren bei unserem Modell der Sentiment-Analyse messen wir hier pro Sekunde eines Films die Anzahl der gesprochenen Wörter, sodass sich je Film eine Zeitreihe ergibt. Als Features der Zeitreihen extrahieren wir ebenfalls Mittelwert und Quartilswerte zur Erfassung der Verteilung der Dialogtempo-Werte, sowie die Rate der Mittelwertdurchgänge jeder Zeitreihe und Nulldurchgangsraten der ersten und zweiten Ableitung zur Abschätzung von periodischen Eigenschaften.

Ähnlichkeitsberechnung

Für alle Untertitel werden anhand der oben genannten Feature-Modelle entsprechende Ergebnisvektoren berechnet (vgl. Abb. 1). Ähnlichkeiten bzw. Distanzen werden pro Modell separat berechnet. Für das *bag of words*-Modell verwenden wir die Cosinus-Ähnlichkeit als Metrik, für alle anderen Modelle die Cosinus-Delta-Metrik, die der Cosinus-Ähnlichkeit auf standardisierten Feature-Werten (*z-Scores*) entspricht und auch häufig in der Stilometrie Verwendung findet. Ein Gesamtähnlichkeitswert zwischen zwei Filmen, wie er in *SubRosa* letztendlich ablesbar ist, wird berechnet als der gewichtete Mittelwert der Ähnlichkeitswerte aus den einzelnen Modellen. Darüber hinaus ist eine spezifische Gewichtung (jeweils 0 - 100%) der einzelnen Features über das Interface des Webtools *SubRosa* möglich.

Feature-Modell	Anzahl der Dimensionen
Bag of words / tf-idf	4952
Sentiment Analyse	6
Stoppwort-Verteilung	87
POS-Trigramme	417
Statistische Maße	10
Dialogtempo	6

Abbildung 1: Anzahl der Dimensionen je Feature-Modell.

Ergebnisse in SubRosa

Wie eingangs beschrieben versteht sich *SubRosa* als exploratives Tool um die Auswirkung unterschiedlicher Features auf die Ähnlichkeitsberechnungen zwischen Untertiteln zu untersuchen. Zur besseren Illustration der Möglichkeiten des Tools zeigt Abb. 2 die grafische Benutzeroberfläche von *SubRosa* mit einer Darstellung ähnlicher Filme zum Film „Alien (1979)“. Auf der linken Seite zu sehen sind die unterschiedlichen Feature-Modelle und deren Gewichtung, die sich jeweils auf die Ergebnisdarstellung auswirken. Die Ergebnisse der Ähnlichkeitsberechnungen zwischen den Filmen werden in einem Graphen visualisiert, in dem jeder Knoten einen Film darstellt und die Länge der Kante zwischen jeweils zwei Filmen näherungsweise proportional zum Quadrat der zwischen ihnen berechneten Distanz ist.

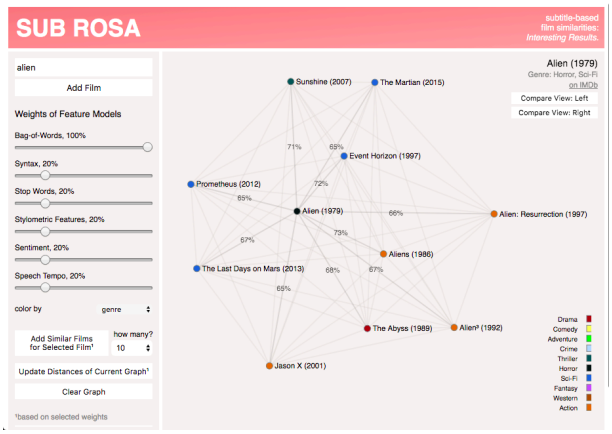

Abbildung 2: Ähnlichkeitsnetzwerk für Alien (1979) in SubRosa.

In Detailansichten (vgl. Abb. 3) für jedes Feature-Modell lassen sich darüber hinaus für jeden einzelnen Film seine extrahierten Feature-Daten analysieren und mit denen anderer Filme vergleichen.

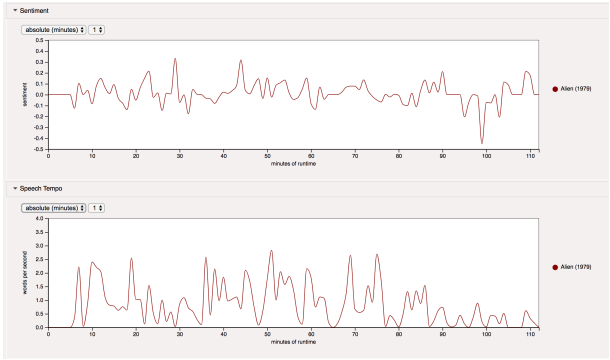

Abbildung 3: Detailsicht der einzelnen Feature-Modelle für Alien (1979), hier für die beispielhaften Features „Sentiment Analyse“ und „Dialogtempo“.

Um einen Überblick zu allgemeinen Ähnlichkeitsmustern im Sinne von Cluster-Bildung innerhalb unseres Korpus an Untertiteln zu erlangen, haben wir zudem den hochdimensionalen Vektorraum jedes Modells mithilfe einer SVD (singular value decomposition) reduziert und die Ergebnisse mittels t-SNE (t-distributed stochastic neighbor embedding) in einem zweidimensionalen Raum als Punkte visualisiert, die entsprechend der Filmgenres eingefärbt sind. Beispielhaft zeigen sich bei der Visualisierung einer gewichteten Kombination aller Modelle (50% Bag-of-Words-Modell, andere Modelle je 10%; siehe Abb. 4) interpretierbare Cluster von Filmen bestimmter Genres, am deutlichsten im Falle von Horror- und Comedy-Filmen. Bei näherer Betrachtung zeigen sich zudem Cluster von Filmen, die sich zwar im Genre stark unterscheiden, jedoch durch ein gemeinsames Setting oder Thema verbunden sind (wie z.B. Weltraum, Western, Schifffahrt, Sport, …).

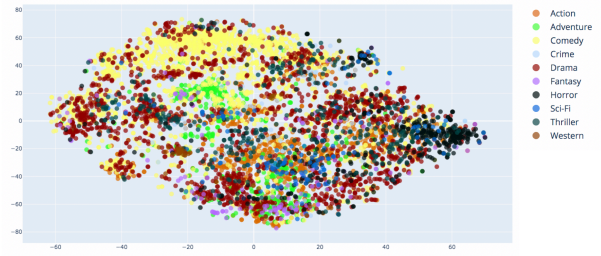

Abbildung 4: Gewichtete Kombination aller Feature-Modelle und 2D-Projektion mittels SVD und t-SNE.

Weiterhin lässt sich zeigen, dass die meisten Features nicht miteinander korrelieren, d.h. Filme die bspw. anhand des Features „Sentiment“ ähnlich sind, können sich erheblich unterscheiden was etwa das Dialogtempo angeht (vgl. Abb. 5).

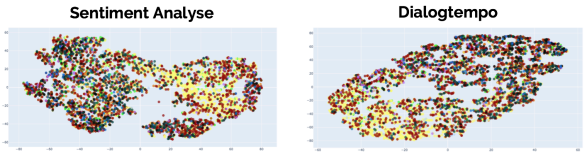

Abbildung 5: Die 2D-Projektion der Untertitel mittels SVD und t-SNE anhand der Features “Sentiment Analyse” und “Dialogtempo” zeigt sehr unterschiedliche Cluster und lässt darauf schließen, dass diese beiden Merkmale nicht korrelieren. Dies gilt im Übrigen auch für die meisten der anderen Features; die entsprechenden Diagramme finden sich online über plot.ly[3].

Die Unterschiedlichkeit der verschiedenen Features lässt sich auch gut anhand beispielhafter Analysen illustrieren. So zeigt sich etwa, dass bei der Suche nach ähnlichen

Filmen zu “The Room” (2003) für jedes einzelne Feature bei den Top 5 der als ähnlich identifizierten Ergebnisse jeweils ganz unterschiedliche Filme herauskommen (vgl. Abb. 6). Einzig “Ruby Sparks” (2012) findet sich sowohl bei “Syntax” als auch bei “Sentiment” wieder. Der Film “The Disaster Artist”, der dokumentationsartig die Entstehungsgeschichte des Klassikers “The Room” schildert (und damit einen unmittelbaren inhaltlichen Bezug hat), kommt interessanterweise nur bei der *bag of words*-Methode in den Top 5 der Ergebnismenge vor. Es zeigt sich also, dass ein multifaktorieller Vergleich von Filmen anhand unterschiedlicher, dialog-basierter Features, nicht zielführend ist, sondern vielmehr unterschiedliche Merkmale unterschiedliche Ähnlichkeitsaspekte kodieren. Im nächsten Schritt planen wir eine systematische Korrelationsanalyse der unterschiedlichen Features, um gemeinsam auftretende Phänomene und Muster für spezifische Filmgenres etc. identifizieren zu können.

Methode	Unsortierte Top 5 Ergebnismenge
Bag of Words	American Reunion (2012), Bridesmaids (2011), The Disaster Artist (2017), This is Where I Leave You (2014), Funny People (2009)
Syntax (=POS-Trigramme)	Ruby Sparks (2012), Shame (2011), Addicted (2014), Pretty in Pink (1986), The Edge of Seventeen (2016)
Stop Words	Belle de Jour (1967), Jamon Jamon (1992), Frankie and Johnny (1991), Sweet November (2001), Some Kind of Wonderful (1987)
Stylometric Features (=Statistische Maße)	Stranger Than Paradise (1984), The Butterfly Effect 3: Revolutions (2009), Shallow Grave (1984), Cellular (2004), The Forgotten (2004)
Sentiment	Willy Wonka and the Chocolate Factory (1971), Night and the City (1950), Bee Movie (2007), Stuck on You (2003), Ruby Sparks (2012)
Speech Tempo	The Hangover Part III (2013), Terminal (2018), The Hateful Eight (2015), The Man Who Wasn't There (2001), Life Itself (2018)

Abbildung 6: Unterschiede in der Ergebnismenge verschiedener Feature-Konfigurationen für „The Room“ (2003).

Fazit und Ausblick

Im hier vorgestellten Projekt dokumentieren wir aktuelle Experimente zur Identifikation von Ähnlichkeitsbeziehungen zwischen Film-Untertiteln auf Basis ganz unterschiedlicher Features, die künftig für quantitative Stil- und Genreanalyse von Filmen herangezogen werden können. *SubRosa* versteht sich zunächst als experimentelle Plattform, die es erlaubt interaktiv unterschiedliche Feature-Kombinationen für unterschiedliche Filme bzw. Fragestellungen zu erproben. Als Verbesserung auf technischer Ebene planen wir die Integration eines größeren Korpus[4] (Lison & Tiedemann, 2016), welches systematischer validiert und korrigiert wurde als es bei unserem aktuellen Testkorpus der Fall ist.

Darüber hinaus soll über eine systematische Evaluation eine Feature-Selektion und optimale Gewichtung erfolgen. Geplant ist hierzu eine Evaluation gegen eine *ground truth* auf Basis bestehender Ähnlichkeitsverbindungen, bspw. über die Empfehlungen via *collaborative filtering* bei Amazon oder über den frei verfügbaren Datensatz *MovieLens*.[5] Offen ist dabei die Frage, ob Ähnlichkeitsbewertungen auf Basis audio-visueller Features grundsätzlich mit Ähnlichkeitsbewertungen auf Dialogebene korrelieren, oder die verschriftlichte Dialogebene ggf. als isolierte Ebene betrachtet werden muss. Wir planen deshalb weitere Fallstudien mithilfe von *SubRosa*, die zusammen mit Film- und Sprachwissenschaftlern durchgeführt werden sollen.

Fußnoten

1. OpenSubtitles: https://www.opensubtitles.org/de
2. IMDb: https://www.imdb.com/
3. Feature-Visualisierungen: https://chart-studio.plot.ly/~bbrause/#/
4. OpenSubtitles 2018-Korpus: http://opus.nlpl.eu/OpenSubtitles2018.php
5. MovieLens Dataset: https://movielens.org/

Bibliographie

Aggarwal, C. C. (2001): On k-anonymity and the curse of dimensionality. In: Proc. 31st International Conference on Very Large Data Bases (VLDB), S. 901–909. ACM, 2005.

Argamon, S. / Shimoni, A. R. / Koppel, M. (2003): Automatically categorizing written texts by author gender. In: Literary and Linguistic Computing, Vol. 17, Nr. 4, S. 401– 412.

Bednarek, M. (to appear 2020): The Sydney Corpus of Television Dialogue: Designing and building a corpus of dialogue from US TV series. Corpora 15/1. Pre-Print-Version hier verfügbar: https://www.monikabednarek.com/wp-content/uploads/2019/09/Designing-and-building-a-corpus-of-US-TV-dialogue_Academia.pdf

Blackstock, A. / Spitz, M. (2008): Classifying movie scripts by genre with a MEMM using NLP-based features. M.Sc. Kurs Natural Language Processing, stud. Projektbericht, Juni 2008. Stanford University.

Bougiatiotis, K. / Giannakopoulos, T. (2017): Multimodal content representation and similarity ranking of movies. Pre-Print-Version hier verfügbar: https://arxiv.org/pdf/1702.04815.pdf

Burghardt, M. / Kao, M. / Wolff, C. (2016): Beyond Shot Lengths – Using Language Data and Color Information as Additional Parameters for Quantitative Movie Analysis. In Book of Abstracts of the International Digital Humanities Conference (DH).

Burghardt, M. / Meyer, S. / Schmidtbauer, S. / Molz, J. (2019): “The Bard meets the Doctor” – Computergestützte Identifikation intertextueller Shakespearebezüge in der Science Fiction-Serie Dr. Who. In Book of Abstracts, DHd 2019.

Schmidt, B. (15.9.2014): Screen time! Published on http://sappingattention.blogspot.com/2014/09/screen-time.html (letzter Zugrifff am 24.9.2019)

Hutto, C. J. / Gilbert, E. (2014): VADER: A parsimonious rule-based model for sentiment analysis of social media text. In: International Conference on Weblogs and Social Media.

Johnson, W. (1944): Studies in language behavior: I. A program of research. In: Psychological Monographs, Vol. 56, S. 1-15.

Kozloff, S. (2000): Overhearing Film Dialogue. University of California Press.

Santini, M. (2004): A shallow approach to syntactic feature extraction for genre classification. In Proceedings of the 7th Annual Colloquium for the UK Special Interest Group for Computational Linguistics (CLUK 2004).

Shannon, C. (1948): A mathematical theory of communication. In: The Bell System Technical Journal, Vol. 27, S. 379–423, 623–656, Juli und October 1948.

Taylor, A. / Tilton, L. (2019): Distant viewing: analyzing large visual corpora. In Digital Scholarship in the Humanities, 2019. Published by Oxford University Press on behalf of EADH.

Torruella, J. / Capsada, R. (2013): Lexical statistics and topological structures: A measure of lexical richness. In: *Procedia - Social and Behavioral Sciences*, Vol. 95, S. 447-454.

Manning, C. / Raghavan, P. / Schütze, H. (2008): Introduction to Information Retrieval. Cambridge University Press.

Müller M. / Volk M. (2013): Statistical Machine Translation of Subtitles: From OpenSubtitles to TED. In: Gurevych I., Biemann C., Zesch T. (eds) Language Processing and Knowledge in the Web. Lecture Notes in Computer Science, vol 8105. Springer, Berlin, Heidelberg.

Nessel, J. / Cimpa, B. (2011): The MovieOracle-content based movie recommendations. In: Proceedings of the 2011 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology, S. 361-364.

Lison, P. / Tiedemann, J. (2016): OpenSubtitles2016: Extracting Large Parallel Corpora from Movie and TV Subtitles. Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC), p. 923-929, European Language Resources Association.

Tiedemann , J. (2016): Finding Alternative Translations in a Large Corpus of Movie Subtitles. In Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC), p. 3518–3522, European Language Resources Association.