Der Einsatz von Distant Reading auf einem Korpus deutschsprachiger Songtexte

Thomas Schmidt, Marlene Bauer, Florian Habler, Hannes Heuberger, Florian Pilsl, Christian Wolff

Lehrstuhl für Medieninformatik, Universität Regensburg, Deutschland

thomas.schmidt@ur.de
marlene.bauer@stud.uni-regensburg.de
florian.habler@stud.uni-regensburg.de
hannes.heuberger@stud.uni-regensburg.de
florian.pilsl@stud.uni-regensburg.de
christian.wolff@ur.de

DHd 2020, Paderborn, Deutschland
März, 2020

Stichwörter: Songtexte, Lyrics, Distant Reading, Sentiment Analysis, Topic Modeling

Zusammenfassung/Abstract. Wir präsentieren die ersten Ergebnisse eines Projekts zur Exploration des Einsatzes von computergestützter Textanalyse und Distant Reading auf einem Korpus deutschsprachiger Songtexte. Der Fokus liegt dabei momentan vor allem auf der Identifikation genrespezifischer Unterschiede für die Genres Pop, Rap, Rock und Schlager. Zu diesem Zweck wurde ein Korpus bestehend aus 4636 Songtexten einiger der bekanntesten Genrevertreter seit den 60er Jahren über die Plattform *LyricWiki* akquiriert. Es werden erste punktuelle Ergebnisse bezüglich Wortfrequenzanalysen, Sentiment Analysis und Topic Modeling präsentiert und diskutiert. Die Wortverteilungen weisen eine homogene Verteilung von in allen Genres auftretenden Konzepten auf, lediglich Rap grenzt sich stärker ab. Ähnliches zeigt sich für die Methoden der Sentiment Analysis und des Topic Modeling. Auch hier werden Unterschiede bezüglich der Verwendung sentiment-beladener Wörter und der Konstitution von Topics insbesondere bezüglich des Genres Rap deutlich.

Please cite as:

Schmidt, T., Bauer, M., Habler, F., Heuberger, H., Pilsl, F. & Wolff, C. (2020). Der Einsatz von Distant Reading auf einem Korpus deutschsprachiger Songtexte. In *DHd 2020 Spielräume: Digital Humanities zwischen Modellierung und Interpretation. Konferenzabstracts (pp. 296-300)*. Paderborn, Germany.

Link zu den Konferenzabstracts: https://zenodo.org/record/3666690#.Xz-cfZMzbUI
Link zum Artikel in den Konferenzabstracts:
https://zenodo.org/record/3666690/preview/2020_DHd_BookOfAbstracts-web.pdf#page=298

Einleitung

Die Idee des Distant Reading (Moretti, 2002) ist davon geprägt, durch den Einsatz von Methoden der computergestützten Textanalyse und Textvisualisierung große Mengen an Literatur zu explorieren, um Einsichten zu gewinnen, die mit herkömmlichen Methoden nicht möglich sind. Der Einsatz von Distant Reading wird dabei mittlerweile auch außerhalb der Literaturwissenschaften untersucht wie z.B. in den

Religionswissenschaften (Pfahler et al., 2018). Im folgenden Beitrag wird ein Projekt vorgestellt, in dem der Einsatz und Nutzen von Distant Reading in ersten Analysen auf einer größeren Menge deutschsprachiger Songtexte exploriert wird. Ziel des Projekts ist es, mittels Distant Reading Unterschiede in gängigen Genres populärer Musik herauszukristallisieren.

Verwandte Arbeiten

Im Bereich des Text Mining wird die Analyse von Songtexten vor allem im Kontext von Retrieval- und Recommender-Aufgaben betrieben. Ziel ist meist die automatische Klassifikation und Vorhersage verschiedener Kategorien, z.B. dem Genre (Fell & Sporleder, 2014; De Sousa et al., 2016). Außerhalb dieses Arbeitsgebiets findet man in Bereichen der Kultur- und Literaturwissenschaften sowie der Psychologie Studien mit Songtexten als Untersuchungsgegenstand (Cole, 1971; Kuhn, 1999). Forschungsinteressen umfassen dabei Analysen spezifischen Musikern (*Beatles*, West & Martindale, 1996; Whissel, 1996, *Bob Dylan*, Whissel, 2008; Körner, 2012), Epochen (Pettijohn & Sacco, 2009), Emotionen (Napier & Shamir, 2018) oder Erfolg (Riedemann, 2012). Im Bereich der computergestützten Korpus-Analyse findet man vereinzelt Projekte für den englischsprachigen Bereich. Dabei werden beispielsweise quantitative und qualitative Methoden verknüpft, um Stil und historische Eigenheiten zu analysieren (Werner, 2012), Annotations- und Akquisemöglichkeiten von Korpora exploriert (Kreyer & Mukherjee, 2009) oder N-Gramme untersucht (Nishina, 2017). Die Analyse von deutschsprachigen Texten ist jedoch bislang selten und findet vor allem im Bereich von regionalem Rap statt (Hess-Lüttich, 2009) sowie eher qualitativ und hermeneutisch (Stiegler, 2009).

Korpus-Erstellung

Als Plattform für die Akquise der Songtexte wurde *LyricWiki*[1] gewählt. Ausgehend von aktuellen Umfragen zu den populärsten Genres in Deutschland[2] werden die folgenden vier Genres betrachtet: *Pop, Rock, Schlager* und *Rap/Hip Hop*. Für die Auswahl der Songs wurden manuell durch Analyse der deutschen Charts seit den 60er Jahren eine angemessene Anzahl der wichtigsten deutschsprachigen Genre-Vertreter aufgestellt. Dieser Schritt ist (auch) subjektiv geprägt, der Fokus auf berühmte und „typische“ Vertreter der einzelnen Genres erlaubt jedoch trotzdem erste Analysen. Kritisch sei jedoch anzumerken, dass die Grenzen der Genres für einzelne Interpreten und Songs nicht immer eindeutig sind, insbesondere was Rock, Pop und Schlager betrifft. Wir haben versucht, für das vorliegende Korpus eine Auswahl mit möglichst eindeutigen Zuordnungen zu treffen.[3]

Für jeden gewählten Interpreten wurden über ein Skript alle Songtexte mit Metadaten von LyricWiki akquiriert. Die Akquise des Korpus wurde mittels eines frei verfügbaren angepassten ruby-Skripts durchgeführt[4].

Abbildung 1 illustriert Eckdaten zum Gesamtkorpus und den Künstlern. In der Spalte „Bekannte Vertreter“ werden einige Künstler beispielhaft angegeben.

[1] https://lyrics.fandom.com/wiki/LyricWiki
[2] https://de.statista.com/statistik/daten/studie/171224/umfrage/beliebteste-musikrichtungen/
[3] Das Korpus ist erhältlich auf Anfrage und über GitHub:
https://github.com/lauchblatt/GermanSongLyricsCorpus
[4] https://gist.github.com/siavashs/3556469

Genre	Künstler	Albums	Songs	Tokens	Bekannte Vertreter
Pop	22	96	1132	302614	Nena, Rosenstolz, Herbert Grönemeyer
Rap	33	129	1558	864925	Die fantastischen Vier, Samy Deluxe, Sido
Rock	20	126	1312	320751	Udo Lindenberg, Die Ärzte, Rammstein
Schlager	16	83	634	147833	Peter Maffay, Wolfgang Petry, Helene Fischer
Gesamt	91	434	4636	1636123	

Abbildung 1: Korpus-Zusammensetzung

Abbildung 2 zeigt die Songverteilung im zeitlichen Verlauf und Genre-Kontext auf.

	Pop	**Rap**	**Rock**	**Schlager**	**All**
60er	-	-	-	10	10
70er	10	-	51	41	102
80er	101	-	132	85	318
90er	98	108	194	141	541
00er	459	694	632	264	2049
10er	463	756	303	94	1616
All	1132	1558	1312	634	4636

Abbildung 2: Genre und zeitlicher Verlauf des Korpus

Im Bereich des Preprocessing wurden Stoppwörter entfernt und alle Wörter zu Normalisierungszwecken in Kleinschreibung gebracht.

Methoden und Ergebnisse

Für die allgemeine Textanalyse und das Topic Modeling wurden alle Analysen mittels *R* und unterschiedlichen Bibliotheken wie dem *NLP*[5]- und *topicmodels*-package[6] durchgeführt. Die Sentiment Analysis wurde mit *Python* und *SentiWS* (Remus et al., 2010) implementiert.

Allgemeine Textanalyse

Die Repetition von besonders bedeutenden Wörtern ist ein gängiges Stilmittel bei der Gestaltung von Songtexten. Aus diesem Grund betrachten wir die Analyse der häufigsten Wörter von Songtexten als besonders aufschlussreich. Die folgenden Bilder (Abbildung 3-6) illustrieren die 10 häufigsten Wörter (Most Frequently Used Words; MFWs) der einzelnen Genres.

[5] https://cran.r-project.org/web/packages/NLP/index.html
[6] https://cran.r-project.org/web/packages/topicmodels/index.html

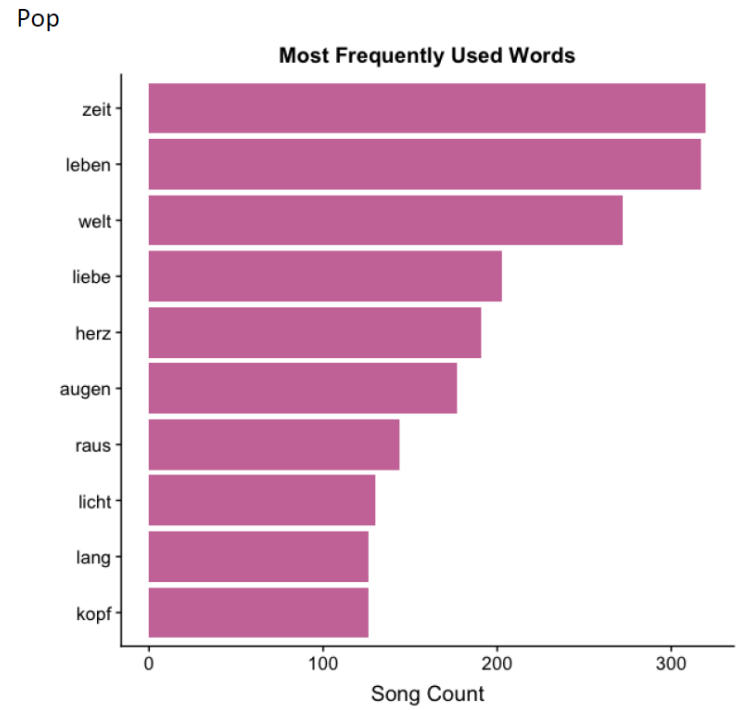

Abbildung 3: MFWs für Pop

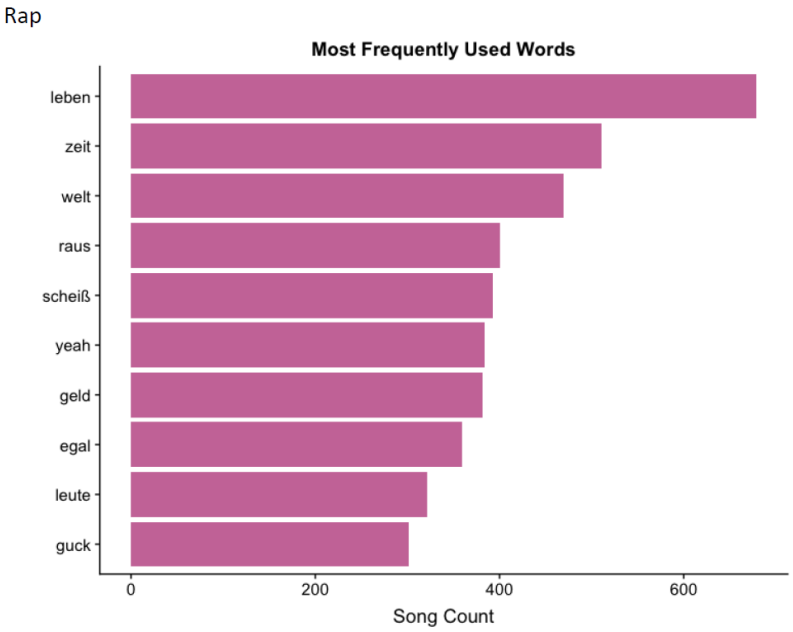

Abbildung 4: MFWs für Rap

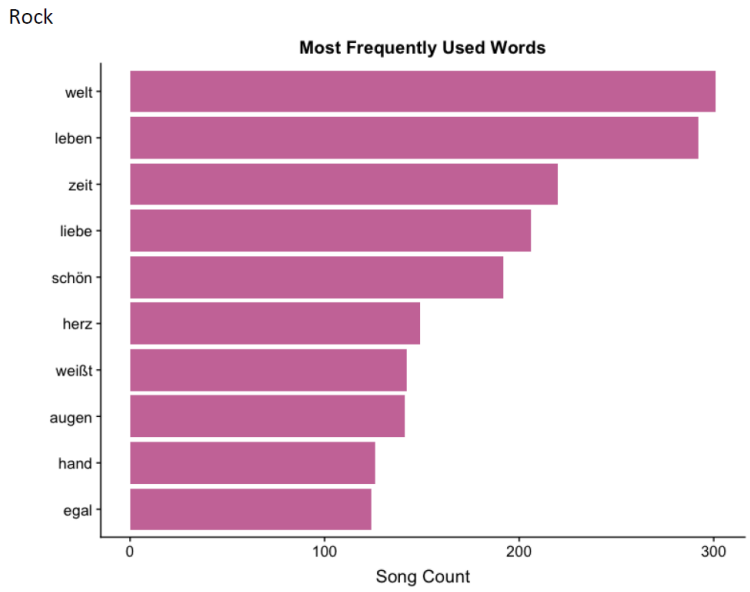

Abbildung 5: MFWs für Rock

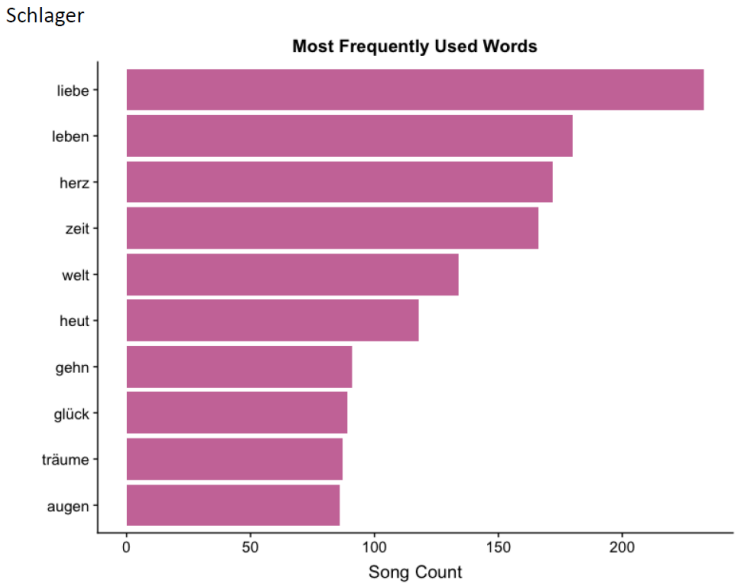

Abbildung 6: MFWs für Schlager

Man erkennt, dass es drei Wörter gibt, die in allen vier Genres gleichmäßig stark vertreten sind: „Welt“, „Leben“ und „Zeit“. Diese Konzepte sind demnach konsistenter Inhalt deutschsprachiger Liedtexte unabhängig vom Genre. Die größte Differenzierung zeigen die Genres Rap, in dem Terme der Umgangs- und Jugendsprache enthalten sind, aber auch thematische Schwerpunkte deutlich werden („Geld“) sowie das Genre Schlager, das vor allem von emotionalen Termen wie „Liebe“, „Herz“ oder „Glück“ dominiert wird.

Sentiment Analysis

Sentiment Analysis ist die Methodik zur computergestützten Analyse von Sentiments in Texten, also ob und in welchem Ausmaß Wörter eines Textes eher positiv oder negativ konnotiert ist (Liu, 2016). In den Digital Humanities werden häufig lexikonbasierte Methoden zur Bestimmung von Sentiment-Werten eingesetzt (Mohammad, 2011; Nalisnick & Baird, 2013). Dabei wird durch Summenbildung von

Sentiment-Werten von Wörtern die Gesamtpolarität einer Texteinheit ermittelt. Wir verwenden dabei das etablierte Sentiment-Lexikon *SentiWS* (Remus et al., 2010). Abbildung 7 illustriert einige Ergebnisse:

Genre	Häufigste positive Wörter (Häufigkeit)	Häufigste negative Wörter (Häufigkeit)	Gesamt-Polarität	Gesamt-Polarität normalisiert an der Anzahl der Tokens
Pop	liebe (628) schön (255) lieben (116) rein (198)	schwer (186) kurz (108) feuer (104) kleine (87) fallen (85)	-1373.681	-0.004539
Rap	liebe (627) rein (399) lieber (377) schön (257) reich (207)	scheiß (718) scheiße (427) hart (317) alter (271) schwer (232)	-363.398	-0.0004201
Rock	liebe (579) schön (349) lieber (210) lieb (150) rein (146)	schwer (144) scheiß (115) feuer (112) kurz (108) kalt (108)	-424.503	-0.0013234
Schlager	liebe (547) nah (115) lieb (108) lieben (104) schön (102)	feuer (114) arm (78) schwer (78) wein (55) fehlt (50)	-344.082	-0.0023275

Abbildung 7: Ergebnisse – Sentiment Analysis

Man erkennt, dass für alle 4 Genres insbesondere Varianten von Liebe einen erheblichen Beitrag zur positivien Polarität leisten. Rap grenzt sich deutlich mit für das Genre typischen Themen ab, ausgedrückt durch Wörter wie „reich“ und mit Slang („hart“, „alter“). Alle Genres weisen insgesamt auf eine negative Polarität hin. Entgegen der naiven Intuition sind die Genres „Rap“ und „Rock“ dabei noch am positivsten (gemessen an den normalisierten Werten) bewertet. Erste Analysen machen jedoch auch Probleme der lexikonbasierten Sentiment-Analyse deutlich. Die Wörter „wein“ (weinen) und „feuer“ (das Feuer) sind in SentiWS als negativ markiert, haben aber in unseren Texten oft eher positive Konnotationen. Bei dem Wort „wein“ dann, wenn dieses durch die Normalisierung von „der Wein“ hergeleitet wird. In zukünftigen Arbeiten wollen wir mit einem domänenspezifischen Lexikon arbeiten, das für die jeweilige Anwendungsdomäne optimiert ist.

Topic Modeling

Topic Modeling ist eine Methode, um den Anteil verschiedener Themen in Dokumenten zu analysieren. Ein Thema ist dabei ein selbst definiertes Label für eine Liste von Wörtern, die besonders häufig zusammen auftreten. Als Algorithmus wurde Latent Dirichlet Allocation (LDA) gewählt (Blei et al., 2003). Das Topic Modeling wurde separat für die einzelnen Genres durchgeführt, um Unterschiede und Gemeinsamkeiten zu untersuchen. Wir sind momentan noch am Anfang der Analyse der einzelnen Topics, aber neben Differenzen werden auch Topics gefunden, die ähnliche Konzepte widerspiegeln. Folgende Visualisierungen geben die Wortlisten wider, die wir jeweils als das Topic „Liebe“ in den

einzelnen Genres benannt haben. Die Wortgröße gibt die Häufigkeit des Wortes im jeweiligen Sub-Korpus wider (Abbildung 8).

Abbildung 8: Wortlisten für das Topic „Liebe“

Auffällig ist, dass insbesondere bei Rap familiäre Begriffe wie „Mama“, „Vater“ oder auch „Bruder“ Bestandteil des Topics sind, was traditionellerweise ein häufiger Schwerpunkt im Rap-Genre ist.

Ausblick

In unseren zukünftigen Arbeiten wollen wir insbesondere das Korpus systematisch vergrößern und verbessern. Momentane Probleme sind z.B. die Ungleichverteilung in der Menge bezüglich der Genres aber auch ein Fokus auf eher aktuelle Künstler. Wenngleich wir schon erste Eigenheiten der Genres feststellen konnten, wollen wir Methoden wie Sentiment Analysis und Topic Modeling noch weiter explorieren, indem wir beispielsweise die Varianz der Sentiments untersuchen. Des Weiteren wollen wir unsere Arbeit aber auch auf andere Textanalyse-Möglichkeiten wie Kollokationsprofile von Keywords, Named Entity Recognition und Stilometrie ausweiten. Durch die Zusammenarbeit mit Musik- und Literaturwissenschaftlern wollen wir in Zukunft auch explorieren, welche weiteren Forschungsfragen mit Hilfe größerer Korpora und Distant Reading-Methoden beantwortet werden können.

Bibliographie

Blei, David M. / Andrew, Y. Ng / Michael, I. Jordan (2003): "Latent dirichlet allocation", in *Journal of machine Learning research* 3: 993-1022.

Cole, Richard R. (1971): "Top songs in the sixties: A content analysis of popular lyrics", in: *American Behavioral Scientist* 14 (3): 389-400.

De Sousa, Jefferson Martins / Eanes Torres, Pereira / Luciana Ribeiro, Veloso (2016): "A robust music genre classification approach for global and regional music datasets evaluation", in: *IEEE International Conference on Digital Signal Processing (DSP).*

Fell, Michael / Caroline Sporleder (2014): "Lyrics-based analysis and classification of music", in: *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics.*

Hess-Lüttich, Ernest WB. (2009): "Rap-Rhetorik. Eine semiolinguistische Analyse schweizerischer rap-lyrics", in: *Ars Semeiotica 32.*

Körner, Stefan (2012): "Bob, Pop, Bibel-die Spuren der Bibel in den Songtexten Bob Dylans." *Pastoraltheologie* 101 (12): 503-521.

Kreyer, Rolf / Joybrato Mukherjee (2007): "The style of pop song lyrics: A corpus-linguistic pilot study." in: *Anglia-Zeitschrift für englische Philologie* 125 (1): 31-58.

Kuhn, Elisabeth D. (1999): "'I just want to make love to you'-Seductive strategies in blues lyrics", in: *Journal of pragmatics* 31 (4): 525-534.

Liu, Bing (2016): *Sentiment analysis: Mining opinions, sentiments, and emotions.* New York: Cambridge University Press.

Mohammad, Saif (2011): "From once upon a time to happily ever after: Tracking emotions in novels and fairy tales.", in: *Proceedings of the 5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities* 105-114.

Moretti, Franco (2002): "Conjectures on World Literature" in: *New Left Review* Jan / Feb: 54–68.

Napier, Kathleen / Lior, Shamir (2018): "Quantitative Sentiment Analysis of Lyrics in Popular Music", in: *Journal of Popular Music Studies* 30 (4): 161-176.

Nalisnick, Eric T. / Baird, Henry S. (2013): "Character-to-character sentiment analysis in shakespeare's plays.", in*: Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics* 479–483.

Nishina, Yasunori (2017): "A study of pop songs based on the billboard corpus." in: *Int. J. Lang. Linguist* 4 (2): 125-134.

Pettijohn, Terry F. / Donald F. Sacco Jr. (2009): "The language of lyrics: An analysis of popular Billboard songs across conditions of social and economic threat", in: *Journal of Language and Social Psychology* 28(3): 297-311.

Pfahler, Lukas / Elwert, Frederik / Tabti, Samira / Morik, Katharina / Krech, Volker (2018): "Versuche zum distant reading religiöser Online-Foren": in *Book of Abstracts, DHd 2018.*

Remus, Robert / Quasthoff, Uwe / Gerhard, Heyer (2010): "SentiWS-A Publicly Available German-language Resource for Sentiment Analysis.", in: *LREC:* 1168-1171.

Riedemann, Frank (2012): "Computergestützte Analyse und Hit-Songwriting.", in: *Black box pop. Analysen populärer Musik*: 43-56.

Stiegler, Christian (2009): *Nur ein Wort.* Dissertation, Universität Wien.

Werner, Valentin (2012): "Love is all around: A corpus-based study of pop lyrics." in: *Corpora* 7 (1): 19–50.

West, Alan / Colin Martindale (1996): "Creative trends in the content of Beatles lyrics" *Popular Music & Society* 20 (4): 103-125.

Whissell, Cynthia (1996): "Traditional and emotional stylometric analysis of the songs of Beatles Paul McCartney and John Lennon", in: *Computers and the Humanities* 30 (3): 257-265.

Whissell, Cynthia (2008): "Emotional fluctuations in Bob Dylan's lyrics measured by the Dictionary of Affect accompany events and phases in his life", in*: Psychological reports* 102 (2): 469-483.