MULTIMODALER BEDEUTUNGSTRANSFER VOM TEXT ZUM BILD. GRANULARE BILDKLASSIFIKATION DURCH VERTEILUNGSSEMANTIK.

Simon Donig
Lehrstuhl für Digital Humanities
Universität Passau, Deutschland
simon.donig@uni-passau.de

Maria Christoforaki
Lehrstuhl für Data Science
Institut für Informatik
Universität St.Gallen, Schweiz
maria.christoforaki@unisg.ch

Bernhard Bermeitinger
Lehrstuhl für Data Science
Institut für Informatik
Universität St.Gallen, Schweiz
bernhard.bermeitinger@unisg.ch

Siegfried Handschuh
Lehrstuhl für Data Science
Institut für Informatik
Universität St.Gallen, Schweiz
siegfried.handschuh@unisg.ch

Dezember 2019

1 Einleitend

In den letzten Jahren hat die Verwendung von Bildklassifizierungsverfahren wie neuronalen Netzwerken auch im Bereich der historischen Bildwissenschaften und der *Heritage Informatics* weite Verbreitung gefunden (Lang und Ommer 2018). Diese Verfahren stehen dabei vor einer Reihe von Herausforderungen, darunter dem Umgang mit den vergleichsweise kleinen Datenmengen sowie zugleich hochdimensionalen Datenräumen in den digitalen Geisteswissenschaften. Meist bilden diese Methoden die Klassifizierung auf einen vergleichsweise flachen Raum ab. Dieser flache Zugang verliert im Bemühen um ontologische Eindeutigkeit eine Reihe von relevanten Dimensionen, darunter taxonomische, mereologische und assoziative Beziehungen zwischen den Klassen beziehungsweise dem nicht formalisierten Kontext. Eine in (Donig, Bermeitinger u. a. 2019) vorgeschlagene Lösung, diese Beziehungen wieder in den Prozess der Klassifizierung zurückzubringen, ist, sich die größere Ausdruckskraft von textbasierten Modellen zunutze zu machen, um die Fähigkeiten visueller Klassifikatoren zu erweitern.

Dabei wird ein *Convolutional Neural Network* (CNN) genutzt, dessen Ausgabe im Trainingsprozess, anders als herkömmlich, nicht auf einer Serie flacher Textlabel be-

ruht, sondern auf einer Serie von Vektoren. Diese Vektoren resultieren aus einem *Distributional Semantic Model* (DSM), welches aus einem Domäne-Textkorpus generiert wird. Ein DSM ist ein multidimensionaler Vektorraum, in dem Wörter als Vektoren abgebildet werden (Lenci 2018).

Das durchgeführte Experiment beruht auf der Kollation von zwei Korpora: Einem textbasierten und einem visuellen. Mit dem Textkorpus wird zunächst ein DSM erzeugt und diesem dann eine Auswahlliste von Zielwörter zugeführt (die funktional den Annotationslabeln der Bilder entspricht). Als Ergebnis erhalten wir Vektoren, die mit diesen Wörtern korrespondieren und mit denen die Bilder annotiert werden. Mit diesen Vektorannotationen wird ein neuronales Netzwerk trainiert, das anschließend dem Klassifikator unbekanntes Bildmaterial identifizieren soll. Als Ergebnis dieses Klassifikationsprozesses erhalten wir einen Vektor, der mithilfe des DSMs in natürlichsprachige Wörter zurückgewandelt wird. Da wir nach reicheren Repräsentationen im Zuge dieses Vorgangs suchen, wählen wir die fünf nächsten Nachbarn aus (*Top-5 Nearest Neighbors*). Als Ähnlichkeitsmaß legen wir die Kosinusähnlichkeit zwischen dem vorhergesagtem Vektor und jenem Vektor zugrunde, der dem ursprünglich dem Bild von uns zugewiesenen Textlabel entspricht. Wir gehen davon aus, dass ein Bild korrekt klassifiziert wurde, wenn das Goldlabel unter diesen Top-5 erscheint.

Darüber hinaus vergleichen wir die Ergebnisse des vorgeschlagenen Klassifizierungsverfahrens mit einem herkömmlichen Verfahren auf der Grundlage flacher Label unter Verwendung desselben CNNs, das für das Vektor-Experiment genutzt wurde. Wir können zeigen, dass das Vektor-Verfahren (bezogen auf die Metriken) ebenso effizient und in einigen Aspekten sogar besser ist.

2 Aufbau des Experiments

Das Experiment beruht auf je einem Bild- und Textkorpus aus dem Bereich Sachkulturforschung mit einem Fokus auf klassizistische Artefakte.

2.1 Textkorpus

Das Textkorpus besteht aus 44 Quellen, die unter einer freien, permissiven Lizenz verfügbar sind, und umfasst englischsprachige Fachpublikationen zu Mobiliar und Raumkunst, die von der Jahrhundertwende bis zur Mitte des 20. Jahrhunderts erschienen sind. Das Textkorpus wurde in mehreren Schritten gereinigt und aufbereitet: Zum einen wurden Standard-NLP-Verfahren angewandt, darunter Tokenisierung, Satz- und Worttrennung, die Normalisierung von Zahlenwerten und die Erkennung von benannten Entitäten (*Named Entity Recognition, NER*). Da wir retrodigitalisiertes Material aus verschiedenen Quellen nutzen, implementieren wir manuelle Korrekturen für die häufigsten der vorkommenden Fehler (etwa Ligaturen wie `II`, die als `U` durch das OCR-Verfahren fehlinterpretiert werden). Eine weitere Ebene der Vorverarbeitung besteht aus inhaltsbezogenen Augmentierungen. Insbesondere normalisieren wir zusammengesetzte Wörter und Synonyme gemäß einer spezifizierten Liste, die anhand einer Ontologie, der *Neoclassica-Ontologie* (Donig, Christoforaki und Handschuh 2016) zusammengestellt sind. Dies resultiert in einem Korpus von insgesamt 3 067 237 Wörtern

aus 107 518 Wortgrundformen. Das DSM wird von uns mithilfe des *Indra*-Frameworks (Sales u. a. 2018) sowie *Gensim* (Řehůřek und Sojka 2010) erzeugt[1].

2.2 Bildkorpus

Das Bildkorpus besteht aus 1231 Ansichten klassizistischer Möbel in deren Gesamtheit, die permissiv lizenziert sind[2] und die sowohl historisches Bildmaterial als auch Fotos aus der modernen Bestandsdokumentation umfassen. Es repräsentiert 28 Klassen.

2.3 Kombiniertes Korpus

Da es sich um ein *Proof-of-Concept*-Experiment handelt, kommt zum Zweck des *Rapid Prototypings* ein an die VGG-Architektur (Krizhevsky, Sutskever und Hinton 2012) angelehntes, „simples“ neuronales Netzwerk zum Einsatz[3]. Die Unabhängigkeit der Trainings- und Evalutationsbeispiele wird durch eine 5-fache Kreuzvalidierung auf 80 % des Korpus garantiert, wovon wiederum 80 % für das Training und 20 % für das Testen benutzt werden. Die übrigen 20 % werden als Evaluationsmenge behandelt. Die reportierten Ergebnisse basieren auf diesen, dem Klassifikator unbekannten 20 %.

Da durch Sammlungspraxis der Gedächtnisinstitutionen (Sammelwürdigkeit, geografischer Schwerpunkt) und Zugänglichkeit des Materials (Lizenzierung und Grad der Sammlungsdigitalisierung) die Verteilung der Artefakte nach Klassen unbalanciert ist (Abb. 1), sind die Klassengewichte dementsprechend angepasst (seltene Klassen werden höher gewichtet als häufig vorkommende Klassen (Johnson und Khoshgoftaar 2019, S. 27)). Um eine Situation zu vermeiden, in der ein *Machine-Learning*-Modell derart an die Trainingsdaten angepasst wird, dass es darin scheitert auf anderen, aber trotzdem ähnlichen Daten, zu generalisieren, wird die übliche *Early-Stopping*-Methode verwendet.

---

[1] Die Erzeugung erfolgt mit einer Vektorgröße von 50, einer Worfenstergröße von 10 und einer minimalen Wortzahl von fünf erzeugt. Als *Word2Vec*-Modell kommt *Skipgram* (Mikolov u. a. 2013) mit *Negative Sampling* zum Einsatz.

[2] Das Korpus wurde aus den Sammlungen des Metropolitan Museum (New York), des Victoria & Albert Museum (London), der Wallace Collection (London) sowie mehreren zeitgenössischen Musterbüchern zusammengestellt.

[3] Das Netzwerk besteht aus drei Convolutional-Blöcken mit jeweils zwei Convolutional-Layer mit 32, 64 und 64 Filter der Größe 3×3. Nach jedem Block folgt ein Maximum-Pooling-Layer der Größe 2×2, sowie ein Dropout-Layer mit einer Dropoutwahrscheinlichkeit von 0,25. Ein Fully-Connected-Block, bestehend aus zwei Fully-Connected-Layers mit jeweils 256 Knoten, steht im Anschluss sowie nochmals ein Dropout-Layer mit 0,5 Dropoutwahrscheinlichkeit. Jeder Convolutional- und Fully-Connected-Layer bis dahin wurde zufällig initialisiert und benutzt *ReLU* als Aktivierungsfunktion. Der letzte Layer ist ein Fully-Connected-Layer mit 50 Ausgabeknoten und benutzt eine lineare Aktivierungsfunktion. Es ist in den beiden Frameworks *Keras* (`https://keras.io`) und *TensorFlow* (`https://tensorflow.org`) implementiert. Beim Training wird der durchschnittliche absolute Fehler durch die Optimierungsfunktion *RMSprop* minimiert.

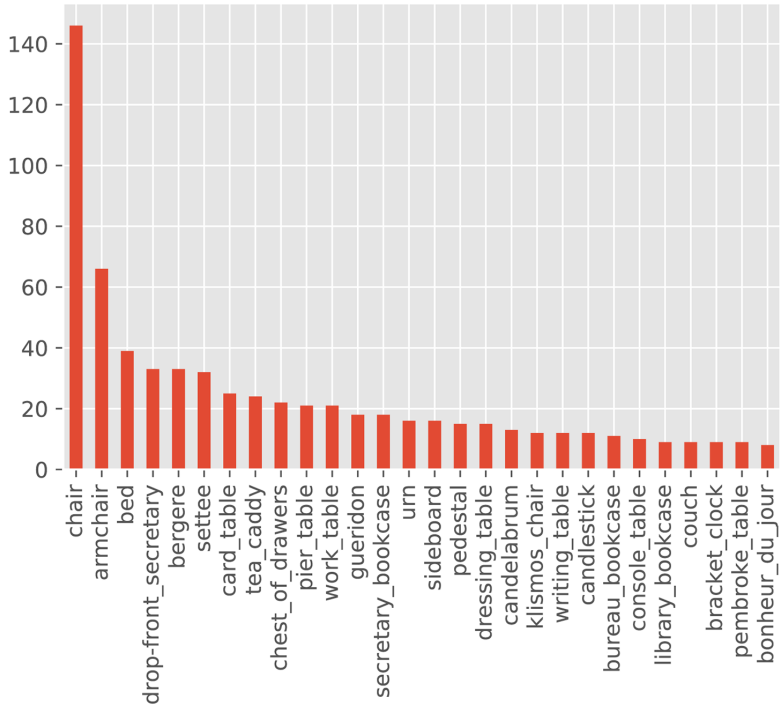

Abbildung 1: Verteilung der Klassen im Bildkorpus.

3 Ergebnisse

Die Top-5-Richtig-Positiv-Rate (*true-positive*) beträgt $0{,}59$. Das bedeutet, dass das Goldlabel in $59\,\%$ der Fälle unter den fünf nächsten Nachbarn erscheint.

Das mathematische Qualitätskriterium gibt für sich genommen jedoch nur einen Teil des Gesamtbilds wieder. Wir führen deshalb zugleich eine qualitative Analyse der Ergebnisse in der Evalutionsteilmenge durch.

Eine Reihe von richtig-positiven Ergebnissen zeigen, dass die Klassifizierung keinesfalls zufällig erfolgt, sondern dass die Top-5-Begriffe tatsächlich jeweils denselben semantischen Nachbarschaften entstammen. Sie drücken eine Reihe von Beziehungen taxonomischer und assoziativer Natur aus. Beispielsweise wird der Roentgen-Schreibtisch aus dem Bestand des V&A in Abb. 2 mit Labeln (in der Reihenfolge) von `dressing_table`, `writing_table` und `work_table` assoziiert (Ankleidetisch, Schreibtisch, Nähtisch). Diese Trias ist schon deshalb sinnhaltig, weil viele dieser Artefakte multifunktional waren und mehrere dieser Funktionen erfüllten. Daneben ähneln auch jene Artefakte, die dezidiert nur einem einzigen Zweck dienten, konstruktiv den jeweils anderen Möbeltypen. Die Nähe der drei Konzepte entsteht also sowohl auf semantischer Ebene (Nähe der Wörter im DSM, die wiederum das Produkt lebensweltlicher Nähe ist), als auch auf einer visuellen Ebene im CNN (visuelle Formähnlichkeit). Ein weiteres Bild desselben Objekts (Abb. 3) zeigt einerseits, dass die Methode in sich

Abbildung 2: Abweichungen bei der Klassifizierung desselben Objekts (Victoria & Albert Museum 2019b).

konsistent ist (die Top-4 sind identisch, obwohl eine andere Perspektive vorliegt) und andererseits, dass auch die visuellen Merkmale innerhalb des CNNs eine Auswirkung auf den Klassifizierungsprozess haben. Da Schreibschränke (*secrétaires à abbatants*) häufig frontal, hochaufrecht und mit einer geöffneten Schreibklappe oder -schublade abgebildet werden, scheint deren Vorkommen im Bild eine Klassifizierung als Sekretär ausgelöst zu haben. Im ersten Bild könnte dagegen die Anwesenheit von Schubladen (*drawers*) zu einer Klassifizierung als Kommode geführt haben, die naheliegenderweise auf semantischer Ebene mit Schubladen assoziiert ist.

Während die Label in den bisher betrachteten Fällen die taxonomischen Beziehungen reflektieren und alle den aus der Ontologie abgeleiteten Zielwörtern entstammen, zeigt Abb. 4, dass das Verfahren auch aus sich selbst, rein datenzentriert Label generieren kann. Die abgebildete Kratervase wurde manuell als Urne (*urn*) annotiert. Die Top-2-Wörter reflektieren demnach auch taxonomischen Beziehungen („urn“, „vase“). Die anderen Konzepte spiegeln dagegen assoziative Beziehungen wider. Das Label `bell` ist ein Artefakt des Reinigungsprozesses, da im Korpus Wörter wie „bell-shaped“ oder „bell-crater“ (mit und ohne Bindestrich) existieren, um diese Art von Artefakten zu beschreiben. „Ovoid“ bezieht sich demgegenüber wohl auf die Eierstabdekoration des oberen Wulstes, die oft mit diesem Adjektiv beschrieben wird. Diese Ornamentik scheint zugleich die Assoziation zur Rosette (*Patera*) mitbedingt zu haben. Auf diese Weise erscheint das Zielwort `patera_element` unter den Top-5, obwohl im Bildkorpus ausschließlich ganze Artefakte, nicht aber deren Dekor annotiert wurden.

Nicht auszuschließen ist hier zudem ein Effekt des visuellen Klassifikators, wie auch Abb. 5 zeigt. Die Fehlklassifikation des Objekts, eines Nähtischchens, führte zu konsistenten Zuschreibungen im Bereich der Sitz- und Liegemöbel. Betrachtet man die

Abbildung 3: Abweichungen bei der Klassifizierung desselben Objekts (Victoria & Albert Museum 2019b).

äußere Form des Artefakts auf einer abstrakteren Ebene, kann man eine visuelle Nähe zu z. B. einem *(Double-) Camel-back Sofa* durchaus nachvollziehen.

4 Vergleich zu einem herkömmlichen CNN

Um die Unterschiede zwischen beiden Zugängen besser abschätzen zu können, führen wir weiter in Vergleichsexperiment durch, bei dem dasselbe CNN[4] wie im vektorbasierten Verfahren für eine herkömmliche Klassifizierung mit flachen Labeln herangezogen wird.

Um die Unterschiede der beiden Zugänge (herkömmliche Klassifizierung mit flachen Labeln/vektorbasiertes Verfahren) besser abschätzen zu können, stellen wir im Folgenden einen Vergleich zentraler Metriken für beide Zugänge vor.

Wie in Tabelle 1 ersichtlich[5], ist nicht nur die Top-1-Treffergenauigkeit im Fall der Klassifizierung mit Vektoren besser, sondern es sind auch die übrigen Metriken vergleichbar gut. Durch den hier vorgeschlagenen Zugang wird also nicht nur die Treffergenauigkeit verbessert, sondern er liefert zugleich eine reichhaltigere Beschreibung des Bildes.

[4]Das verwendete CNN ist identisch mit dem im Vektor-Experiment verwendeten Netzwerk, mit Ausnahme des letzten Layers, der aufgrund der Labelklassifikation nun 28 statt 50 Ausgaben besitzt und die Aktivierungsfunktion *Softmax* statt einer linearen.

[5]Die englischen Begriffe für bessere Verständlichkeit: Treffergenauigkeit (*Accuracy*), Genauigkeit (*Precision*), Trefferquote (*Recall*), Sensitivität (*True-Positive Rate*)

Abbildung 4: Eine Sèvres-Kopie der Medici-Vase stößt die Klassifizierung mit assoziativen Labeln an. (Victoria & Albert Museum 2019a)

Metrik	Vektor	Flache Label
Top-1-Treffergenauigkeit	0,50	0,40
Top-1-Genauigkeit	0,32	0,29
Top-1-Trefferquote	0,25	0,26
Top-1-F1-Maß	0,26	0,25
Top-5-Sensitivität	0,59	0,73
Top-5-Falsch-Positiv-Rate	0,41	0,27

Tabelle 1: Vergleich zentraler Metriken beider Zugänge.

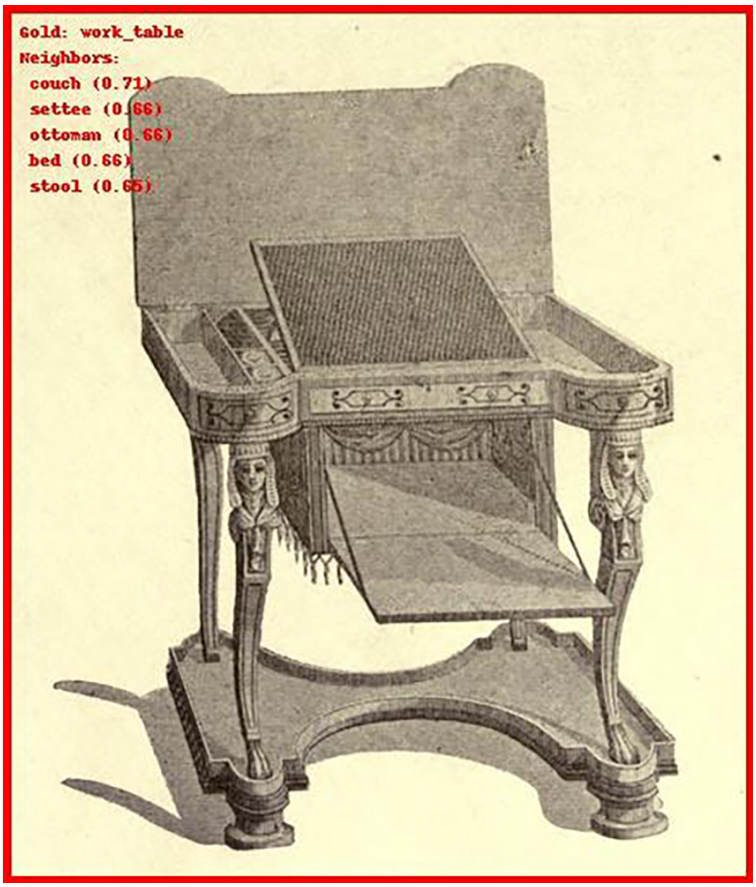

Abbildung 5: Fehlklassifikation eines Nähtischs in ein Wortumfeld aus der Sitzmöbel-Hierarchie. (Sheraton, Hayden und Bell 1910, S. 88)

5 Schlussfolgerungen und Ausblick

In dem hier vorgeschlagenen Beitrag stellen wir ein neues, multimodales Verfahren für die Klassifikation von Bildinhalten vor, das auf der Kombination von NLP-Methoden mit Bildklassifizierungsverfahren beruht. Ziel ist, Objekte nicht alleine nach einem Schema flacher Label, sondern in einer kontextgerechteren Weise zu klassifizieren, wobei dieser Kontext von einschlägigen historischen Domänepublikationen gebildet wird. Dieses Klassifizierungsverfahren bietet einen Zugang zur multidimensionalen Einbettung der Artefakte in die Lebenswelt und deren sprachlicher Widerspiegelung. Dieser Umstand ist von besonderem Nutzen um multifunktionale Objekte zu klassifizieren, ohne dabei auf mehrere Klassifikatoren und einen komplexen Annotationsprozess mit mehreren Labeln zurückgreifen zu müssen. Die Ergebnisse sind ermutigend. Auch mit einem sehr einfachen CNN erreichen wir eine Treffergenauigkeit von $0{,}59$. Als nächsten Schritt möchten wir mit einem komplexeren CNN und einem ausgeweiteten Bildkorpus trainieren (um bekannte Probleme wie *Overfitting* zu reduzieren). Das Ver-

gleichsexperiment mit einem herkömmlichen, auf flachen Labeln beruhenden Zugang zeigt, dass unter Effizienzgesichtspunkten, d. h. im direkten Vergleich der Metriken, unser Verfahren nicht nur vergleichbare Resultate liefert, sondern zugleich auch in einer reichhaltigeren Beschreibung des Bildes resultiert.

Wir werden weiter daran arbeiten, besser zu verstehen, wie ein bestimmtes Textkorpus sich in den Labeln widerspiegelt, die das DSM automatisch zuweist und die nicht Teil der Liste der Zielwörter sind. Ein besseres Verständnis dieser Prozesse scheint insbesondere im Hinblick auf die relativ überschaubaren Textkorpora relevant, die in den Geisteswissenschaften zu spezifischen Themenkomplexen kollationiert werden können. Nicht zuletzt werden wir aus diesem Grund die Nutzung von Thesauri und Wörterbüchern in Betracht ziehen, um Synonymlisten für Zielwörter zu erstellen. In ähnlicher Weise ziehen wir in Betracht, benannte Entitäten zu URIs zusammenzufassen. Das würde uns erlauben spezifische Entitäten (z. B. Werkstätten, Ebenisten und Eigentümer) mit bestimmten Objekten zu assoziieren.

Wir denken, dass dabei der multimodale Zugriff einen besonders effizienten Zugang zu geistes- und kulturwissenschaftlichen Korpora bietet, die, verglichen mit den Korpora anderer Disziplinen in den Natur- und Sozialwissenschaften, klein und Domänerestringiert sind.

Literatur

Donig, Simon, Bernhard Bermeitinger u. a. (März 2019). „Vom Bild Zum Text Und Wieder Zurück“. In: *Digital Humanities: Multimedial & Multimodal - Konferenzabstracts*. DHd2019. Frankfurt am Main, Germany. URL: https://www.researchgate.net/publication/332275547.

Donig, Simon, Maria Christoforaki und Siegfried Handschuh (2016). „Neoclassica - A Multilingual Domain Ontology“. In: *Computational History and Data-Driven Humanities: Second IFIP WG 12.7 International Workshop, CHDDH 2016, Dublin, Ireland, May 25, 2016, Revised Selected Papers*. Hrsg. von Bojan Bozic u. a. Cham: Springer International Publishing, S. 41–53. ISBN: 978-3-319-46224-0. DOI: 10.1007/978-3-319-46224-0_5. URL: https://doi.org/10.1007/978-3-319-46224-0_5.

Johnson, Justin M. und Taghi M. Khoshgoftaar (Dez. 2019). „Survey on Deep Learning with Class Imbalance“. In: *Journal of Big Data* 6.1, S. 27. ISSN: 2196-1115. DOI: 10.1186/s40537-019-0192-5. URL: https://journalofbigdata.springeropen.com/articles/10.1186/s40537-019-0192-5 (besucht am 16. 12. 2019).

Krizhevsky, Alex, Ilya Sutskever und Geoffrey E. Hinton (2012). „ImageNet Classification with Deep Convolutional Neural Networks“. In: *Proceedings of the 25th International Conference on Neural Information Processing Systems - Volume 1* (Lake Tahoe, Nevada). NIPS'12. USA: Curran Associates Inc., S. 1097–1105. URL: http://dl.acm.org/citation.cfm?id=2999134.2999257 (besucht am 16. 12. 2019).

Lang, Sabine und Björn Ommer (1. Dez. 2018). „Attesting Similarity: Supporting the Organization and Study of Art Image Collections with Computer Vision“. In: *Digital Scholarship in the Humanities* 33.4, S. 845–856. ISSN: 2055-7671, 2055-768X. DOI: 10.1093/llc/fqy006. URL: https://academic.oup.com/dsh/article/33/4/845/4964861 (besucht am 16. 12. 2019).

Lenci, Alessandro (14. Jan. 2018). „Distributional Models of Word Meaning“. In: *Annual Review of Linguistics* 4.1, S. 151–171. ISSN: 2333-9683, 2333-9691. DOI: 10.1146/annurev-linguistics-030514-125254. URL: http://www.annualreviews.org/doi/10.1146/annurev-linguistics-030514-125254 (besucht am 16. 12. 2019).

Mikolov, Tomas u. a. (6. Sep. 2013). „Efficient Estimation of Word Representations in Vector Space“. In: arXiv: 1301.3781 [cs]. URL: http://arxiv.org/abs/1301.3781 (besucht am 16. 12. 2019).

Řehůřek, Radim und Petr Sojka (22. Mai 2010). „Software Framework for Topic Modelling with Large Corpora“. In: *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*. Valletta, Malta: ELRA, S. 45–50.

Sales, Juliano Efson u. a. (7.–12. Mai 2018). „Indra: A Word Embedding and Semantic Relatedness Server“. In: *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. Hrsg. von Nicoletta Calzolari (Conference chair) u. a. Miyazaki, Japan: European Language Resources Association (ELRA). ISBN: 979-10-95546-00-9.

Sheraton, Thomas, Arthur Hayden und J. Munro Bell (1910). *The Furniture Designs of Thomas Sheraton*. London: Gibbings and Co., Ltd. URL: https://archive.org/details/furnituredesigns00sheruoft.

Victoria & Albert Museum (2019a). *Vase [Sèvres Copy of the Medici Vase], Paris, 1813. Ascension Number 396-1874*. URL: http://collections.vam.ac.uk/item/O8978/vase-sevres-porcelain-factory/.

– (2019b). *Writing Table, Neuwied, Workshop of David Roentgen ca. 1774-1780. Ascension Number: 1059:1 to 9-1882*. URL: http://collections.vam.ac.uk/item/O117298/writing-table-roentgen-david/.