Universität Regensburg

Philosophische Fakultät III
Sprach- , Literatur- und Kulturwissenschaften
Institut für Information und Medien, Sprache und Kultur (I:IMSK)
Lehrstuhl für Medieninformatik

Praxisseminar
Modul: MEI – M 26.1
SoSe 2015
Leitung: Prof. Dr. Christian Wolff

„Only as good as the Source Material“:

Eine vergleichende Evaluation von OMR-Tools mit der Hoerburger Liedblattsammlung

(vorläufige Fassung)

Katia Buchhop, Florian Fuchs, Miriam Nickl, Thomas Schmidt
Email: Thomas.schmidt@stud.uni-regensburg.de

Inhalt

Abbildungen

Tabellen

1 Einleitung

Im Rahmen dieses Projekts wurde eine Evaluation verschiedener Tools im Bereich der Optischen Musik-Erkennung (OMR) durchgeführt. Dieses Dokument gibt zunächst eine allgemeine Einführung über das Projekt und Optical Music Recognition, anschließend werden Vorgehen und Ergebnisse bezüglich der durchgeführten Evaluation aufgezeigt. Zum Schluss folgt eine kurze Diskussion.

Ein Ziel dieses Projekts ist das Entwickeln eines möglichst effizienten Workflows zur Digitalisierung einer, an der Universität Regensburg archivierten Liedblattsammlung. Eine zentrale Frage ist, inwieweit dieser Digitalisierungsprozess maschinell durchgeführt werden kann, um sowohl Aufwand, als auch die benötigten Arbeitskräfte gering zu halten. Eine Möglichkeit, um gedruckte Noten maschinell einzulesen, bietet die sogenannte Optical Music Recognition (OMR). Diese ist in der Lage, notierte Musik auf eingescannten Bildern zu erkennen und in diversen Formaten wie Music-XML abzuspeichern.

Das Vorgehen von OMR-Tools kann in vier Schritte unterteilt werden. Als erstes werden die Notenlinien identifiziert. Sie bilden die Grundlage für die Erkennung aller nachfolgenden Noten und Symbole. Als nächstes werden die Notenlinien ignoriert, um die Objekte innerhalb dieser zu lokalisieren. Wichtig ist hier zu definieren, wo bestimmte Objekte aufhören und ob sie gegebenenfalls zu anderen Objekten dazugehören. Dieser Fall kann zum Beispiel auftreten, wenn mehrere Noten mit Hilfe eines Balkens verbunden sind. Im nächsten Schritt werden die lokalisierten Objekte einzeln voneinander genau untersucht. Bei Noten wird hier bestimmt, ob es sich um ausgefüllte oder leere Notenköpfe handelt, oder welchen Notenwert der Hals der jeweiligen Note angibt. Als letztes werden die Objekte zueinander in Kontext gesetzt. Hier wird unter anderem untersucht, ob eine Note innerhalb einer Notenzeile ein bestimmtes Vorzeichen zugewiesen bekommen hat (Bainbridge & Bell, 2001).

Noch schwieriger als das Erkennen von gedruckten musikalischen Werken wird es bei handschriftlichen Notationen, worum es sich im Rahmen dieses Projekts handelt. Problematisch ist vor allem die große Variation der verschiedenen ohnehin komplexen Symbole wie Notenschlüssel zwischen den verschiedenen Aufzeichnern. Nach Rebolo et al. (2012) erbringt Optical Music Recognition bei Handschriften wenig aussichtsreiche

Ergebnisse. Der Mehrwert von OMR in Bezug auf die in diesem Projekt behandelte Digitalisierung der vorliegenden Liedblattsammlung wird in diesem Dokument untersucht. Ein weiteres Ziel ist es zu untersuchen welches OMR-Tool am besten zur Erschließung des Datenbestands geeignet ist um begründete Aussagen für das weitere Vorgehen mit der Liedblattsammlung zu machen.

2 Vorgehen

Dieses Kapitel beschreibt den durchgeführten Evaluationsprozess von verschiedenen OMR-Tools anhand eines Test-Korpus aus der Liedblattsammlung. Zunächst wird die Auswahl der Tools und des Test-Korpus erläutert. Anschließend wird das genaue Evaluationsverfahren, sowie die Definition der Metriken beschrieben.

2.1 Tools

Die Auswahl von OMR-Tools ist im Vergleich zur Optical Character Recognition (OCR) sehr begrenzt. Für dieses Projekt wurden drei Tools in Anlehnung an ein bestehendes Evaluationspaper (Bellini, Bruno & Nesi, 2007) gewählt. Generell gibt es vier marktführende kommerzielle Programme: Photoscore[1], SmartScore[2], SharpEye[3] und Capella-Scan[4].

Die Entscheidung für diese Arbeit fiel einerseits auf Photoscore, da dieses als einziges eine manuelle Option für Handschrifterkennung bereit hält. Andererseits wurde SharpEye gewählt, weil es in dem angesprochenen Paper am besten abschneidet. Als drittes Tool wurde Capella-Scan gewählt, da es noch nicht in die im Paper behandelte Evaluation miteinbezogen wurde. Weil SmartScore dort nicht so gut abschneidet wie SharpEye, wird es für die in diesem Dokument beschriebene Evaluation als vernachlässigbar angesehen (Bellini, Bruno & Nesi, 2007). Ein kurzes Antesten von SmartScore hat außerdem keine vielversprechenden Ergebnisse geliefert. Ferner weisen auch einige Web-Quellen auf diese drei Tools als die „besten“ und wichtigsten für das OMR[5].

[1] http://www.sibelius.com/products/photoscore/ultimate.html
[2] http://www.klemm-music.de/musitek/download/index.php
[3] https://www.columbussoft.de/SharpEye.php
[4] http://www.capella.de/de/index.cfm/produkte/capella-scan/info-capella-scan/
[5] http://digitalhumanities.org/answers/topic/what-is-the-best-music-scanning-software

Generell sei zu erwähnen, dass ein objektiver Vergleich der Programme unter anderem wegen der Unterschiedlichkeit der verwendeten Algorithmen kritisch zu sehen ist (Bellini, Bruno & Nesi, 2007). Die Evaluation im Rahmen dieses Projekts bezieht sich ausschließlich auf den hier verwendeten Test-Korpus, der im Folgenden erläutert wird.

Es wurden keine Vollversionen der angesprochenen Programme verwendet, sondern zeitlich beschränkte Demoversionen. Hinsichtlich der Evaluationsergebnisse sollte sich hier jedoch kein Unterschied zu den kostenpflichtigen Varianten ergeben.

2.2 Testkorpus

Die gesamte Liedblattsammlung, welche in der Universität Regensburg archiviert ist, umfasst ca. 20 000 Liedblätter aus dem Bereich der deutschsprachigen Volksmusik. Zu den Liedblättern existieren bereits digitalisierte Metadaten, welche unter anderem Incipits, Sangesort, Herkunft und Titel beinhalten. Diese Metadaten wurden für die Auswahl des Test-Korpus zunächst auf Vollständigkeit geprüft, um einen langen, zusammenhängenden Bereich an Blättern mit möglichst vollständigen Metadaten innerhalb der gesamten Sammlung zu finden. Ein Ziel dieses Projekts ist auch die Zusammenführung der digitalisierten Liedblätter mit den digitalen Metadaten, weswegen es förderlich ist wenn die Auswahl für einen Test-Korpus möglichst viele Metadaten enthält.

Innerhalb des identifizierten Bereichs wurden schließlich ca. 350 Liedblätter händisch, beidseitig eingescannt, aus denen der Testkorpus für die Evaluation der beschriebenen OMR-Tools definiert werden soll. Während des Scanvorgangs machte sich eine große Heterogenität bezüglich der Beschaffenheit der vorliegenden Liedblätter bemerkbar.

Diese Heterogenität offenbarte viele mögliche Probleme für eine effektive Notenerkennung. So sind die Lieder von vielen verschiedenen Leuten niedergeschrieben und die Handschrift weist eine große Variation in der Klarheit auf. Darüber hinaus sind die Textstrophen auf vereinzelten Blättern in die Notenzeilen hineingeschrieben. Es kommt außerdem vor, dass Noten durchgestrichen, oder durch Kommentare ergänzt sind.

Das war ausschlaggebend dafür, dass der Test-Korpus im Fall der OMR-Tool Evaluation zunächst bewusst klein gehalten wurde, da bei einem schlechten Ergebnis auch eine geringere Stichprobe ausreicht, um die maschinellen Notenerfassung innerhalb des

Digitalisierungs-Workflows zu verwerfen. Auch gab es in den Testversionen keine Möglichkeit der maschinellen Ausgabe von Ergebnissen, was dazu führt, dass die Evaluation sehr viel händische Arbeit und Zeit in Anspruch genommen hat. Die Anzahl der Blätter im Testkorpus wurde deshalb aus 20 beschränkt. Für die Auswahl dieser Blätter wurden die gesamten eingescannten Blätter analysiert, mit dem Ziel, möglichst jede unterschiedliche Ausprägung hinsichtlich des Schriftbildes aus den 350 Scans in der Evaluation abzudecken um eine bessere Abbildung der Ergebnisse auf den Gesamt-Korpus zu erreichen.

Diverse Besonderheiten der einzelnen Blätter im Test-Korpus werden in der folgenden Tabelle dargestellt. Die Besonderheiten wurden ausgewählt, da davon ausgegangen wurde, dass die gewählten Gesichtspunkte einen besonderen Einfluss auf das OMR-Ergebnis haben. Inwiefern das zutrifft wird an späterer Stelle beschrieben. Die Unterscheidungen beziehen sich auf Schriftbild, Notenkopf, Notenhals, Notenlinien, Fremdzeichen und Kontrast. Die Erklärungen zu den einzelnen Spalten sind der Tabelle zu entnehmen.

Tabelle 1: Test-Korpus – Besonderheiten Klassifikation

Wert	**Erklärung**
Schriftbild	Dieser Wert bezieht sich auf das allgemeine Schriftbild der Notation. Hier wird berücksichtigt, wie nah die einzelnen Zeichen aneinander geschrieben sind und wie gleichmäßig die Noten gezeichnet wurden.
Notenkopf	Die Größe der Notenköpfe auf dem jeweiligen Notenblatt.
Notenhals	Die Länge der Notenhälse auf dem jeweiligen Notenblatt.
Notenlinien	Dieser Wert bezieht sich lediglich auf den Kontrast der Notenlinien. Einige Blätter haben mit Bleistift gezogene Notenlinien, welche folglich auf dem Notenblatt heller erscheinen.
Fremdzeichen	Der Wert bezieht sich darauf, ob in der Notation Zeichen vorkommen, welche nicht zur eigentlichen Notation gehören. Hierzu zählen Kommentare oder Liedtext, der in die Notenzeilen hineingedruckt ist.
Kontrast	Der allgemeine Kontrast der Notation auf dem Notenblatt.

Sämtliche Werte, der nun folgenden Tabelle wurden lediglich nach einem persönlichen Eindruck erstellt und nicht nach festen Skalen eingeteilt. Es gibt beispielsweise keinen

Längenwert, ab welchem ein Notenhals als kurz oder lang gilt. Auch gibt die Liste die exakten Liedblätter für die Studie an, mit der Signatur als Identifikationsmittel.

Tabelle 2: Testkorpus – Beschreibung

Signatur	Schriftbild	Notenkopf	Notenhals	Notenlinien	Fremdzeichen	Kontrast
A59389	sauber	klein	lang	normal	nein	hoch
A59394	sauber	klein	lang	normal	nein	hoch
A59465	unsauber	klein	kurz	normal	nein	normal
A59906	sauber	klein	kurz	normal	ja	hoch
A60022	sauber	groß	lang	normal	nein	hoch
A60051	unsauber	groß	kurz	normal	ja	hoch
A61630	sauber	klein	kurz	normal	ja	gering
A61815	sauber	groß	lang	normal	ja	hoch
A61816	unsauber	groß	kurz	normal	nein	hoch
A61818	unsauber	groß	lang	hell	nein	normal
A61825	unsauber	groß	kurz	hell	ja	normal
A61826	unsauber	groß	kurz	hell	nein	normal
A61827	unsauber	groß	lang	hell	ja	normal
A61833	unsauber	groß	lang	normal	nein	hoch
A61858	unsauber	groß	lang	normal	ja	normal
A61862	sauber	groß	lang	normal	nein	hoch
A60019	sauber	groß	lang	normal	ja	hoch
A60060	unsauber	groß	kurz	normal	ja	gering
A61852	unsauber	groß	kurz	normal	ja	normal
A61869	unsauber	groß	lang	normal	nein	hoch

Der gesamte Korpus liegt als jpeg-Dateien und als PDFs im Anhang bei.

2.3 Methodik

Der OMR-Prozess bei den zu vergleichenden Tools ist überwiegend gleich. Die Programme akzeptieren als Eingabeformate PDF, JPEG oder TIF. Eine optimierte Kontrasteinstellung für die Scans wird automatisch von den Tools vorgenommen. Photoscore bietet zusätzlich die Option, vor dem OMR-Vorgang auf handschriftliche Noten hinzuweisen. Da in den Fällen von Photoscore, SharpEye und Capella-Scan keine Vollversionen verwendet wurden und die Testversionen nicht die Möglichkeit des Ergebnisexports anbieten, konnte keine maschinelle Auswertung und Gegenüberstellung der Ausgabe erfolgen. Dafür liefern die Programme eine optische Gegenüberstellung der erkannten Musikzeichen mit den originalen Scans. Dadurch ergibt sich die Möglichkeit, händisch die erkannten Zeichen auszuzählen. Hierfür wurden bei den 20 Blättern des Testkorpus im Vorfeld alle relevanten Objekte innerhalb der Notationen analysiert, um deren Gesamtzahl zu ermitteln und sie in verschiedene Kategorien einzuteilen.

Bei Photoscore können freihand Notenlinien definiert werden, anhand derer die Zeichen auf dem Liedblatt interpretiert werden. So kann bei Scans, auf denen die Notenlinien etwas schief sind, die Linienführung für die Interpretation während dem OMR-Vorgang korrigiert werden. Auch hat Photoscore manchmal Notenlinien gar nicht erkannt. Auch dann besteht die Möglichkeit Linien händisch einzufügen. Somit wurde im Fall Photoscore versucht, über diese Funktionalität, immer das Maximum an Erkennungsmöglichkeiten zu erreichen. Vielmals war die händische Anführung von Linien jedoch nicht erfolgreich. Der Output liefert dann ähnlich schlechte Ergebnisse, als wenn gar keine Notenlinien angegeben sind. Diese Methodik wurde gewählt um das maximal beste Ergebnis von Photoscore zu erhalten, auch wenn man feststellen kann, dass der Vorgang der händischen Auszeichnung von Notenlinien nachteilig für den Digitalisierungs-Workflow ist.

Im Rahmen dieser Evaluation wurden nun getrennt voneinander folgende Zeichen behandelt: Noten, Noten mit Sonderzeichen (punktierte Noten, Noten mit Vorzeichen), Pausen, Taktstriche, Sonderzeichen (Violinschlüssel, Bassschlüssel, Tonartangaben). Dabei orientiert man sich an Bellini, Bruno und Nesi (2007), welche noch mehr Kategorien enthalten. Aufgrund der geringeren Komplexität von Volksliedern hat man sich jedoch auf diese Hauptkategorien beschränkt. Die Einteilung in Kategorien soll einerseits zeigen, ob die Programme gegebenenfalls nur mit bestimmten Zeichenarten besondere

Probleme haben, alle anderen Zeichen aber signifikant besser erkennen. Andererseits kann so eine Priorisierung bestimmter Kategorien stattfinden. Es kann beispielsweise angenommen werden, dass die Erkennung der Noten an sich wichtiger ist, als die der Taktstriche. Solche Prioritäten werden im Ergebnis, der in dieser Arbeit durchgeführten Auszählung, berücksichtigt.

Für jedes Zeichen besteht die Möglichkeit, dass dieses komplett richtig, falsch, oder gar nicht erkannt wurde. Außerdem könnte ein Zeichen durch Fehlinterpretation eines anderen Zeichens hinzugefügt worden sein. Die händische Ergebnisauszählung wurde einzeln für jedes Blatt und pro Kategorie in eine Tabelle eingetragen, um anschließend statistische Berechnungen zu ermöglichen.

2.4 Metriken

Für den Vergleich verschiedener Programme im Bereich Optical Music Recognition werden nach (Bellini, Bruno, Nesi & 2007) verschiedene Metriken definiert. Auch orientiert man sich an äquivalenten Studien aus dem OCR (Alexandrov, 2003; Carrasco, 2014; Kanungo, Marton & Bulbol, 1998). Darauf aufbauend soll auch die Evaluation im Rahmen dieses Projekts stattfinden.

Anhand der Ergebnisauszählungen können verschiedene statistische Werte ermittelt werden, mithilfe derer die einzelnen OMR-Programme miteinander verglichen werden können. Nach welchen Gesichtspunkten die Auszählung stattfand und welche statistischen Werte damit erhoben wurden, wird in der folgenden Tabelle dargestellt.

Wert	Berechnung	Erläuterung
Gesamt	Summe	Anzahl der gesamten in einer Kategorie vorkommenden Zeichen
Erkannt	Summe	Anzahl der korrekt erkannten Zeichen
Falsch erkannt	Summe	Anzahl der falsch erkannten Zeichen
Nicht erkannt	Summe	Anzahl der nicht erkannten Zeichen
Hinzugefügt	Summe	Anzahl der fälschlicherweise hinzugefügten Zeichen
Correct	Erkannt / Gesamt	Anteil der korrekt erkannten Zeichen (je größer desto besser)
Confused	Falsch erkannt / Gesamt	Anteil der falsch erkannten Zeichen (je kleiner desto besser)
Lost	Nicht erkannt / Gesamt	Anteil der nicht erkannten Zeichen (je kleiner desto besser)
Spurious	Hinzugefügt / Gesamt	Zeichen, welche fälschlicherweise hinzugefügt wurden als Anteil an der Gesamtzeichenanzahl (je kleiner desto besser)
Error-Rate	(Falsch erkannt + nicht erkannt + hinzugefügt) / Gesamt	Anteil der Summe aus falsch, nicht erkannten, und hinzugefügten Zeichen an der Gesamtzeichenzahl (je kleiner desto besser)
Precision	Erkannt / (Gesamt + hinzugefügt)	Anteil der richtig erkannten Zeichen an der Summe aus Gesamtzeichenzahl und der fälschlicherweise hinzugefügten Zeichen (je größer desto besser)

Für die Auswertung werden die Metriken zunächst für jede Zeichenkategorie einzeln und anschließend in Summe betrachtet.

Für den Spurious-Wert des gesamten Testkorpus wurde explizit eine andere Berechnung als in dem genannten Paper verwendet. Jede Zeichenkategorie hat von sich aus einen eigenen Spurious-Wert, nach (Bellini, Bruno & Nesi, 2007) wurden diese Kategoriewerte zusammengezählt und deren Anteil an der Gesamtzeichenzahl berechnet. Im Rahmen dieses Projekts wurden dagegen in einer erneuten händischen Auszählung nur

jene Zeichen gezählt, welche in Bezug auf den kategorie-übergreifenden Gesamtzeichensatz als tatsächlich hinzugefügt gelten.

Der Grund dafür ist, dass eine vom OMR-Programm als Note interpretierte Pause aus Sicht der Kategorie Note als hinzugefügt gilt, da die Note auf dem Liedblatt nicht existiert, sondern eine Pause ist. Wird zur Berechnung des kategorie-unabhängigen, gesamten Spurious-Wertes jedoch die Summe aus den kategorie-abhängigen Spurious-Werten gewählt, so würde im angesprochenen Fall die Note auch für den Gesamtzeichensatz als zusätzlich hinzugefügtes Zeichen gelten, obwohl in Betrachtung der Gesamtzeichenzahl kein zusätzliches Zeichen hinzugefügt wurde, sondern lediglich ein anderes Zeichen falsch erkannt wurde.

Alle Metriken werden meist prozentual betrachtet, da dies intuitiv zugänglicher ist. Dies hat keinen Einfluss auf die Statistik. Correct In Percent ist eine zentrale Metrik im OMR/OCR und wird auch Accuracy genannt. Eine weitere Metrik die genauer betrachtet wird ist die Error Rate, da diese alle Fehler in einem Maß zusammenfasst und auch die Noise (also Spurious-Anteile) miteinberechnet. In der Literatur gibt es keinen Konsens über die konkrete Interpretation der Metriken. Holley (2009) spricht bei einer Erkennungsrate von über 90% (also Correct In Percent > 90%) von gutem OCR. Eine analoge Interpretation für Error Rate würde gutes OMR bei 10 – 20% sehen. Holley (2009) weist aber auch auf die Grenzen der Interpretation hin, da die Leistung stark von der Qualität des Korpus abhängt.

3 Ergebnisse

Im Folgenden werden die Ergebnisse der Evaluation präsentiert. Die statistische Auswertung wurde mit der Statistik-Software SPSS[6] durchgeführt. Alle genutzten Tabellen und die Ergebnisse befinden sich im Anhang.

Zunächst wird der Test-Korpus deskriptiv beschrieben bevor die Ergebnisse des Tools Photoscore detailliert aufbereitet werden. Auch die Ergebnisse der weiteren Tools werden mit Hilfe deskriptiver Statistik erläutert, dies jedoch deutlich knapper gehalten. Mit Inferenzstatistik wird für die wichtigsten Variablen bewiesen, dass Photoscore die

[6] http://www-01.ibm.com/software/de/analytics/spss/

beste OMR-Leistung bietet. Abschließend wird dies mit einigen Vergleichsgrafiken verdeutlicht. Aufgrund der geringen Zahl an Liedblättern sind alle statistischen Ergebnisse mit Vorsicht zu betrachten.

3.1 Deskriptive Statistik

3.1.1 Testkorpus

Der Testkorpus besteht aus 20 Liedblättern. Eine genaue Beschreibung findet man unter Kapitel 2.1. Folgende Tabelle zeigt die Konstitution des Testkorpus in Bezug auf die vorhandenen musikalischen Zeichen auf:

Tabelle 3: Deskriptive Statistik – Testkorpus

Deskriptive Statistiken

	N	Minimum	Maximum	Mittelwert	Standardabweichung	Varianz
Total General	20	35,00	182,00	89,7000	46,63757	2175,063
Total Notes	20	24,00	127,00	57,2500	30,84575	951,461
Total Notes With Special Characters	20	,00	21,00	4,5000	5,72621	32,789
Total Rests	20	,00	7,00	2,0500	2,08945	4,366
Total Bar Lines	20	1,00	34,00	14,4000	9,82692	96,568
Total Special Characters	20	3,00	30,00	11,5000	7,61923	58,053
Gültige Anzahl (listenweise)	20					

Ein Notenblatt hat im Schnitt ca. 90 Zeichen, davon sind die meisten Zeichen Noten. Die Varianz ist jedoch sehr groß zwischen einem Minimum von 35 und 182 Zeichen. Bezüglich Noten mit Sonderzeichen und Pausen liegen Notenblätter vor, die keines dieser Zeichen enthalten

Folgendes gestapeltes Balkendiagramm zeigt die Verteilung aller Zeichen und die Gesamtzusammensetzung des Korpus auf. Der Korpus beinhaltet dabei insgesamt genau 1794 Zeichen:

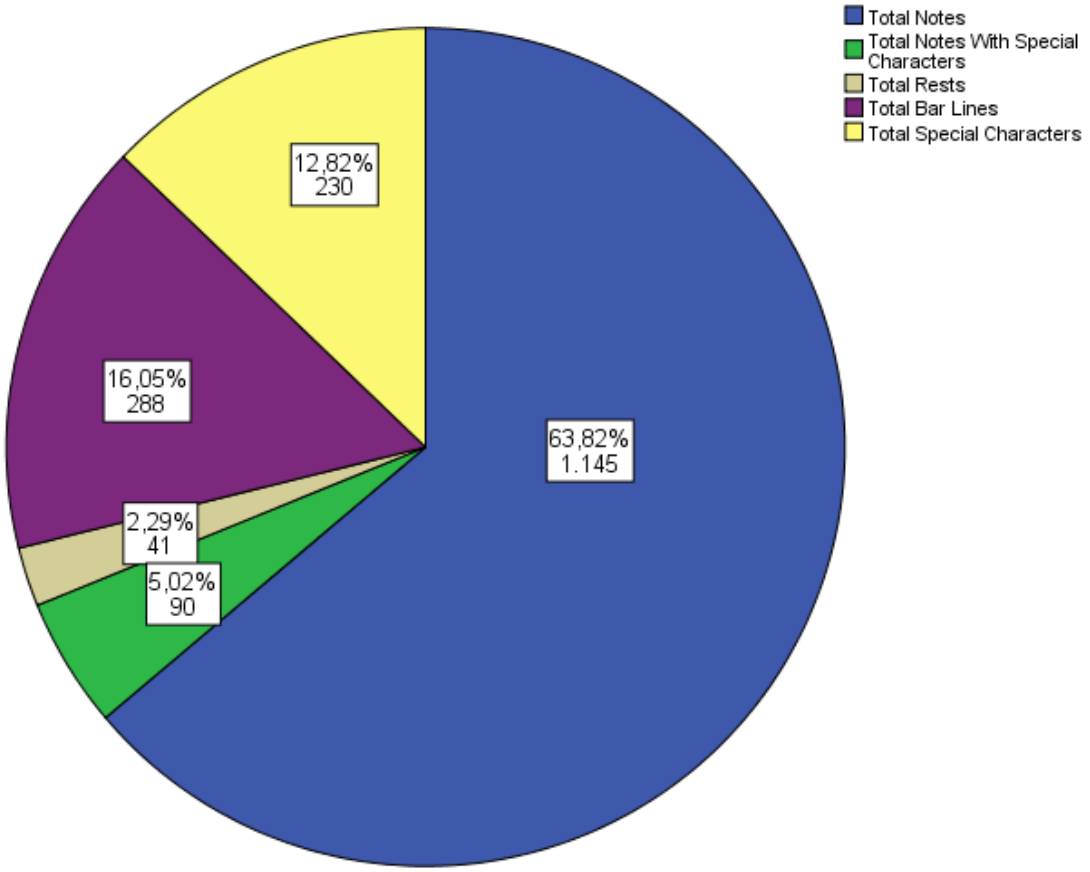


Abbildung 1: Kreisdiagramm – Testkorpus

Im Kreisdiagramm wird ersichtlich, dass Noten den größten Anteil des Korpus ausmachen. Dies bestätigt die Auswahl der Noten-Metriken als zentral für die Performanz-Einschätzung. Mehr als ein Viertel des Korpus besteht noch aus Taktstrichen und Sonderzeichen wie Violinschlüssel und Tonartangaben. Pausen und Noten mit Sonderzeichen stellen nur einen Bruchteil im Gesamtkorpus dar.

3.1.2 Photoscore

Da Photoscore die offensichtlich beste Leistung bei der Erkennung vollbracht hat, werden die Ergebnisse genauer behandelt. Zuerst wird die Leistung auf alle Zeichen gesamt untersucht und daraufhin auf die einzelnen Zeichentypen hin analysiert. Wichtigste Metriken sind dabei Correct In Percent Total, Error Rate Total und die gleichen Maße für die Noten. Diese stellen den Hauptbestandteil für die Zeichenmenge dar und in der Anforderungsanalyse konnte festgestellt werden, dass für wissenschaftliche Arbeiten an dem Korpus, der melodische Verlauf, also die Noten, am wichtigsten sind. Neben diesen

Maßen werden nur die bedeutendsten Befunde noch genauer beschrieben und visualisiert. Alle Ergebnisse lassen sich jedoch im Anhang einsehen.

3.1.2.1 Gesamt (kategorie-unabhängig)

Folgende Tabelle fasst die wichtigsten deskriptiven Daten für alle relevanten Metriken zusammen. Dabei werden die Zeichen kategorie-unabhängig, also gesamt betrachtet:

Tabelle 4: Deskriptive Statistik – Gesamt Photoscore

Deskriptive Statistiken

	N	Minimum	Maximum	Mittelwert	Standardabweichung	Varianz
Total General	20	35,00	182,00	89,7000	46,63757	2175,063
Correct In Percent Total Photoscore	20	,00	75,53	35,9252	28,78305	828,464
Confused In Percent Total Photoscore	20	,00	41,18	17,2203	13,87252	192,447
Lost In Percent Total Photoscore	20	4,26	100,00	46,8545	40,45758	1636,816
Spurious In Percent Total Photoscore	20	,00	114,29	32,9121	35,81986	1283,062
Error Rate Total Photoscore	20	37,23	150,00	96,9869	24,63313	606,791
Precision Total Photoscore	20	,00	66,98	24,5187	20,24449	409,840
Gültige Anzahl (listenweise)	20					

Mit einem Mittelwert von 36% für die Accuracy, hat man eine Erkennungsrate die weit unter erwünschten Grenzwerten liegt wie z.B. 90 % bei Holley (2009). Da das Maximum 75% beträgt wird dieser Betrag für kein Liedblatt erhalten. Die Varianz ist groß und viele Blätter haben eine Erkennung von 0%. Dabei handelt es sich um Blätter bei denen Notenzeilen gar nicht erkannt wurden und das OMR nicht korrekt durchgeführt werden konnte. Dazu mehr in Kapitel 4. Das folgende Histogramm verdeutlicht diese Befunde:

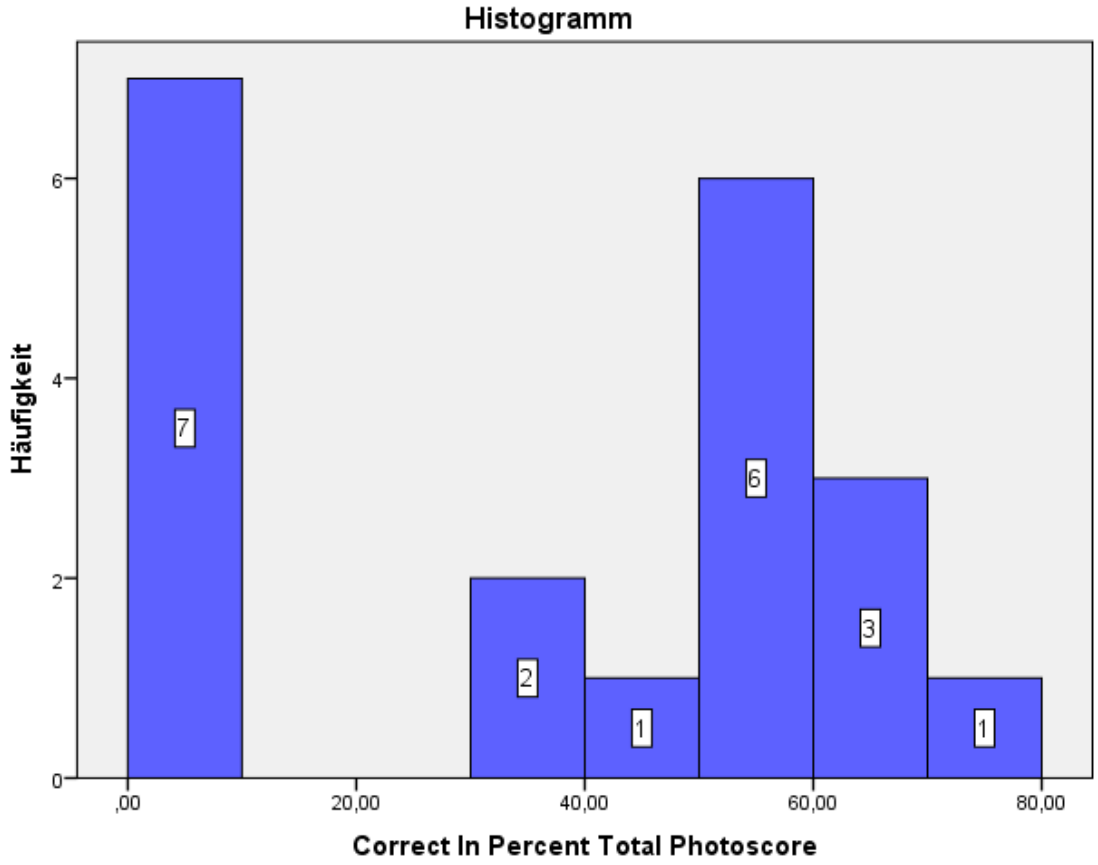


Abbildung 2: Histogramm – Correct In Percent Total Photoscore

Sieben Blätter haben eine Erkennungsrate von 10% oder schlechter, werden also quasi gar nicht erkannt. Folgendes Kreisdiagramm verdeutlicht dabei die Problematik:

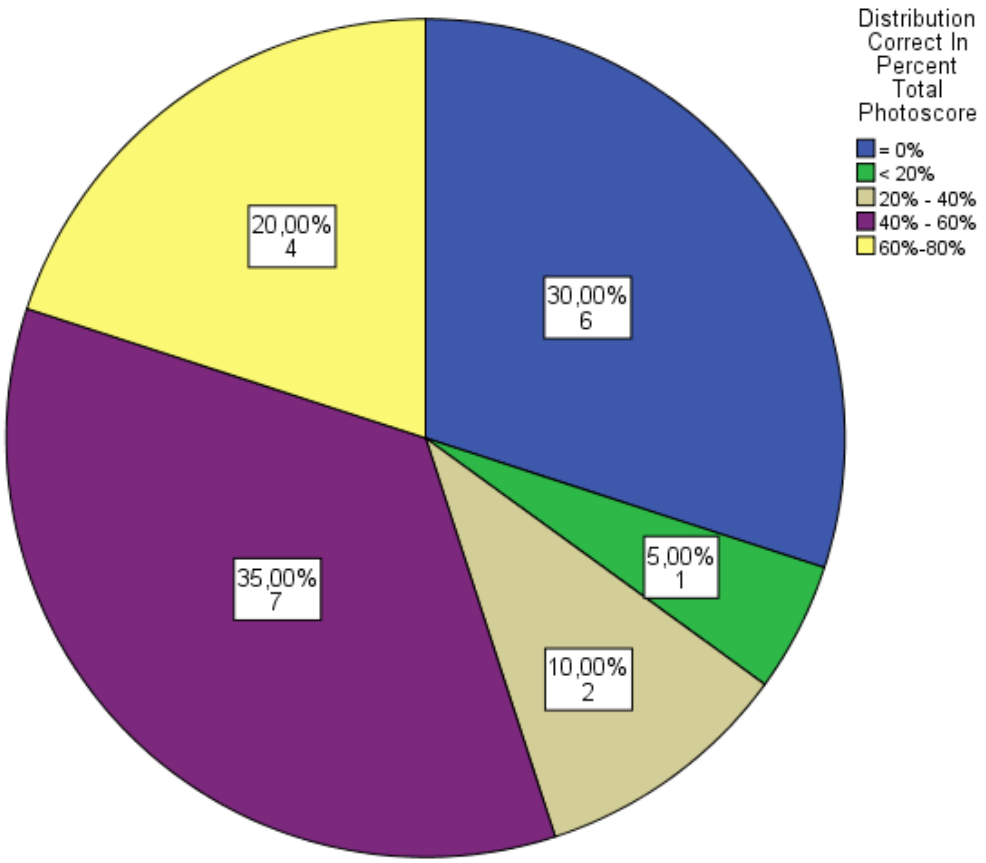


Abbildung 3: Kreisdiagramm – Verteilung Correct In Percent Photoscore

Fast ein Drittel aller Blätter wird demnach gar nicht erkannt, und fast die Hälfte unter 40%. Nur 4 Blätter haben eine Erkennungsrate über 60%.

Untersucht man dabei die Fehlertypen wird bestätigt, dass hauptsächlich der Lost-Typ diese Probleme verursacht. Also Zeichen überhaupt nicht erkannt werden. Mit einem Mittelwert von 47% ist er der am häufigsten auftretende Fehlertyp. Dies bedeutet, dass die Hälfte aller Zeichen nicht erkannt werden. Tatsächlich handelt es sich aber um ganze Notenblätter die nicht erkannt werden und dadurch diesen Fehlertyp statistisch verstärken. Dementsprechend ähnelt das Histogramm auch dem der Accuracy:

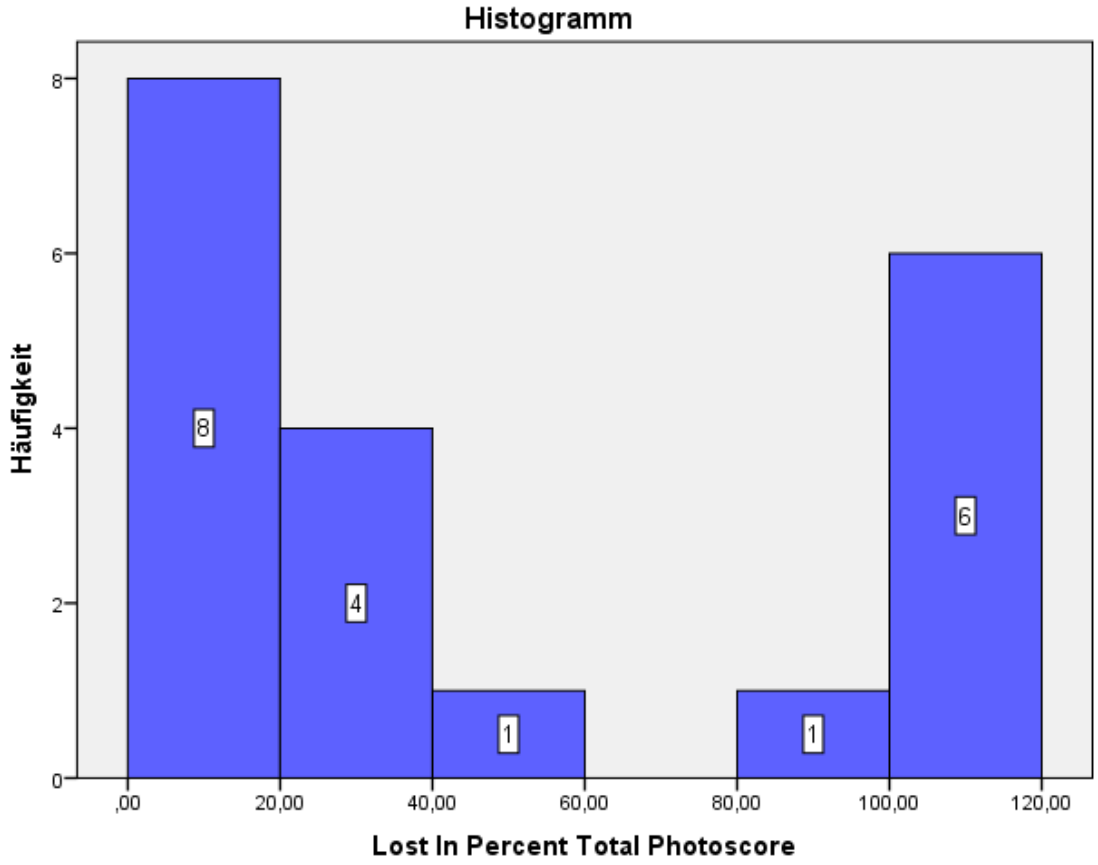


Abbildung 4: Histogramm – Lost In Percent Total Photoscore

Hier wird auch das große Gefälle deutlich. Bei 8 Blättern gehen fast keine oder sehr wenige Zeichen verloren während eben 6 Blätter faktisch gar nicht erkannt werden. Mit einem Mittelwert von 19% sind die Confused-Fehler im Vergleich dazu eher geringfügig.

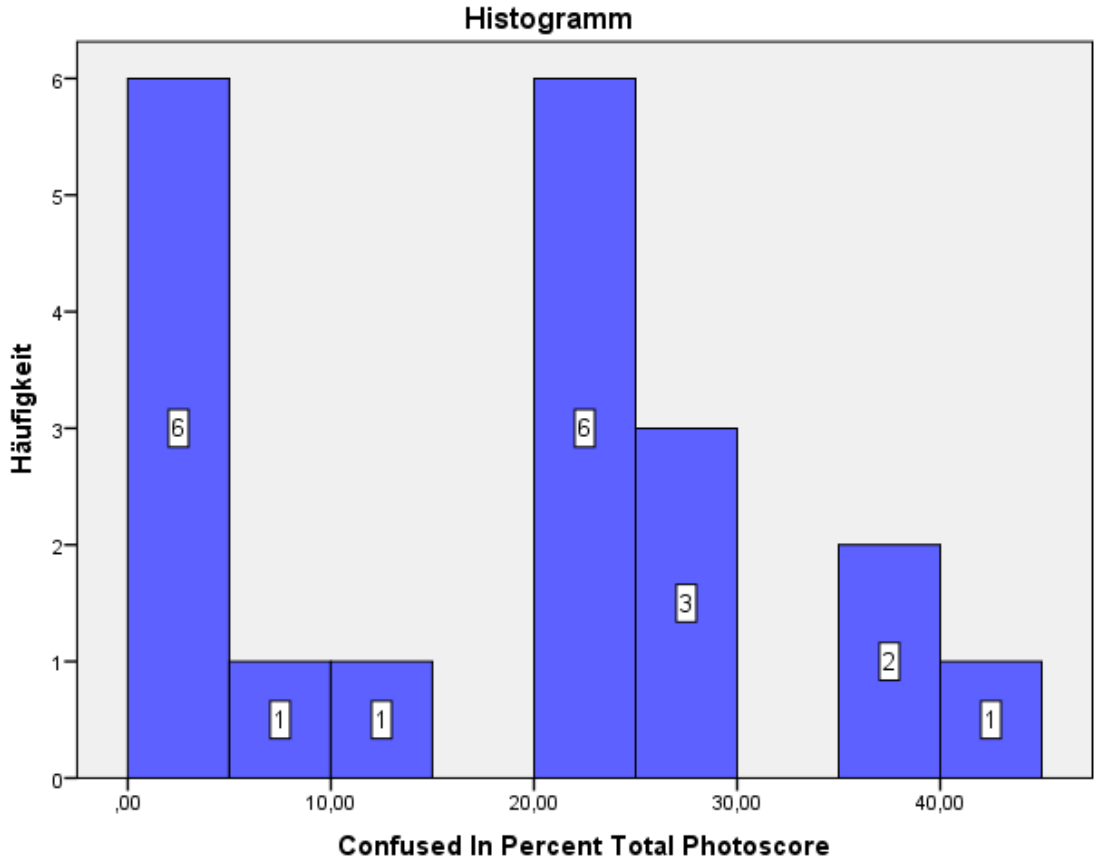


Abbildung 5: Histogramm – Confused In Percent Total Photoscore

Hier ist jedoch zu bedenken, dass aufgrund der hohen Lost-Raten andere Fehler a priori weniger auftreten können. So sind die sechs Blätter mit 0%-Raten bei Confused In Percent eben diejenigen, die vollständig verloren gegangen sind. Bei den anderen Blättern hat man mehrheitlich Confused-Raten von über 20%.

Mit einem Mittelwert von 33% muss man auch konstatieren, dass die Spurious-Werte auf Probleme bei der Noise-Erzeugung weisen. Das Histogramm verhält sich dabei ähnlich wie bei Confused In Percent, folgt aber eher einer gleichmäßigen Verteilung:

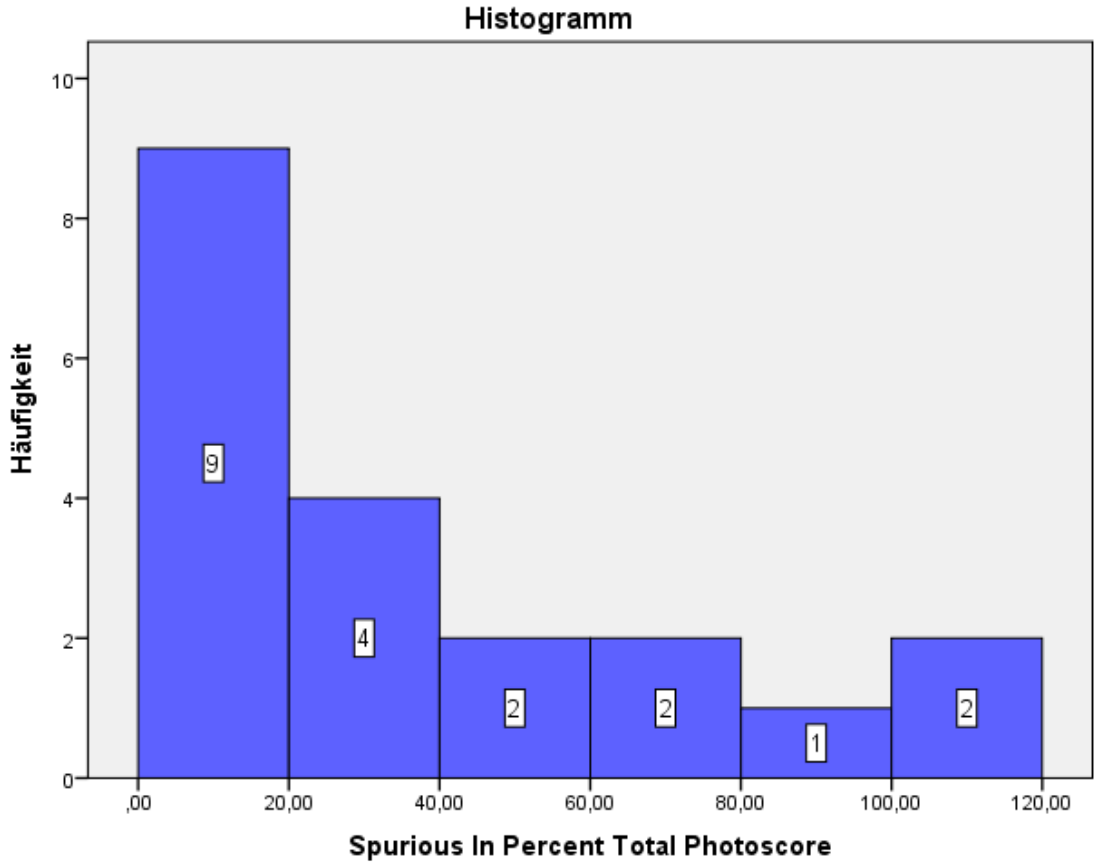


Abbildung 6: Histogramm – Spurious In Percent Total Photoscore

Auch hier liegt eine Verzerrung vor, da bei Blättern die 100% bei der Lost-Rate haben keine Noise erzeugt werden kann. Bei den restlichen Blättern liegen gleichmäßig verteilt Werte zwischen 20% - 120% vor. Das bedeutet, dass in Bezug auf das Liedblatt mehr als ein Viertel an unnötigen Zeichen hinzugefügt wurde. Bei zwei Liedblättern sogar mehr Zeichen als überhaupt im Lied auftreten.

Die Error Rate zeigt, unabhängig von Fehlerabhängigkeiten wie oben, die Fehleranfälligkeit auf, da alle Fehler zusammengefasst werden. Mit einem Mittelwert von 97% ist diese sehr hoch. Dies bedeutet dass in Bezug auf die Anzahl der Zeichen eines Liedblattes im Schnitt fast genauso viele Zeichen im OMR-Output falsch, verloren oder zusätzlich vorliegen. Die Verteilung sieht dabei folgendermaßen aus:

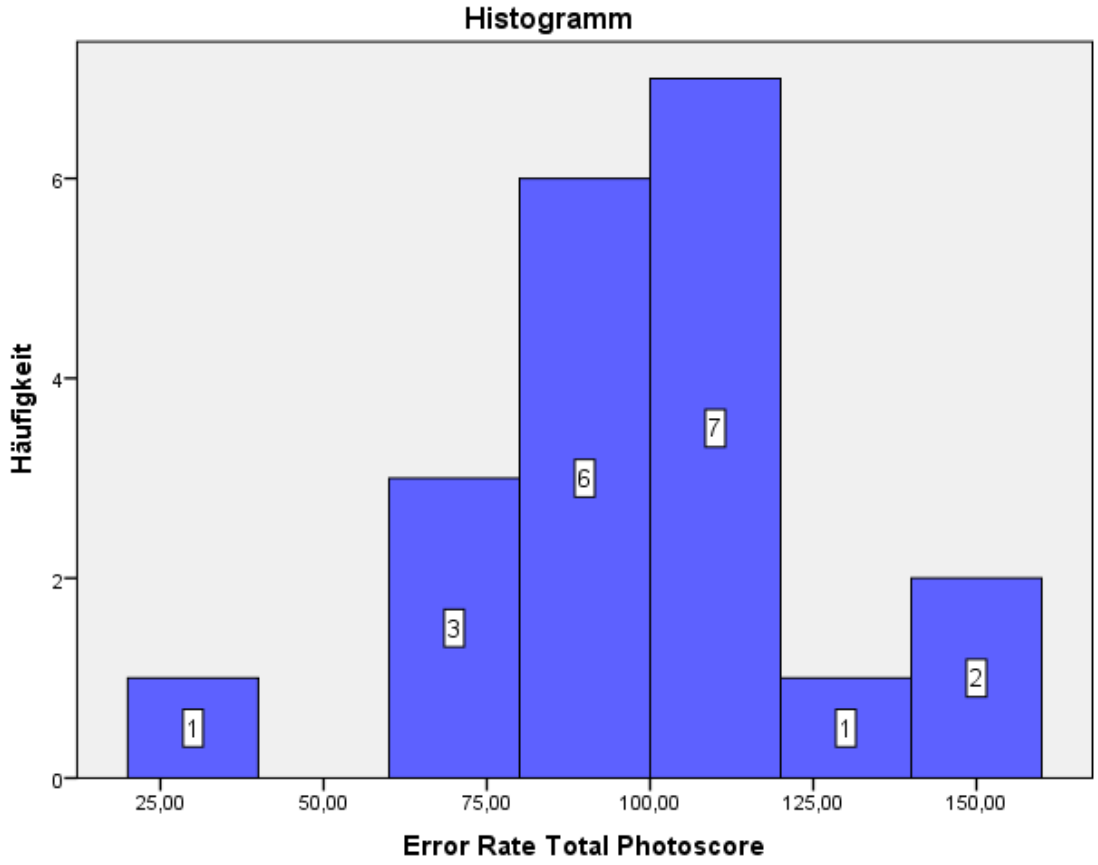


Abbildung 7: Histogramm – Error Rate Total Photoscore

Betrachtet man alle Typen von allen Zeichen im Vergleich, ergibt sich folgendes Kreisdiagramm:

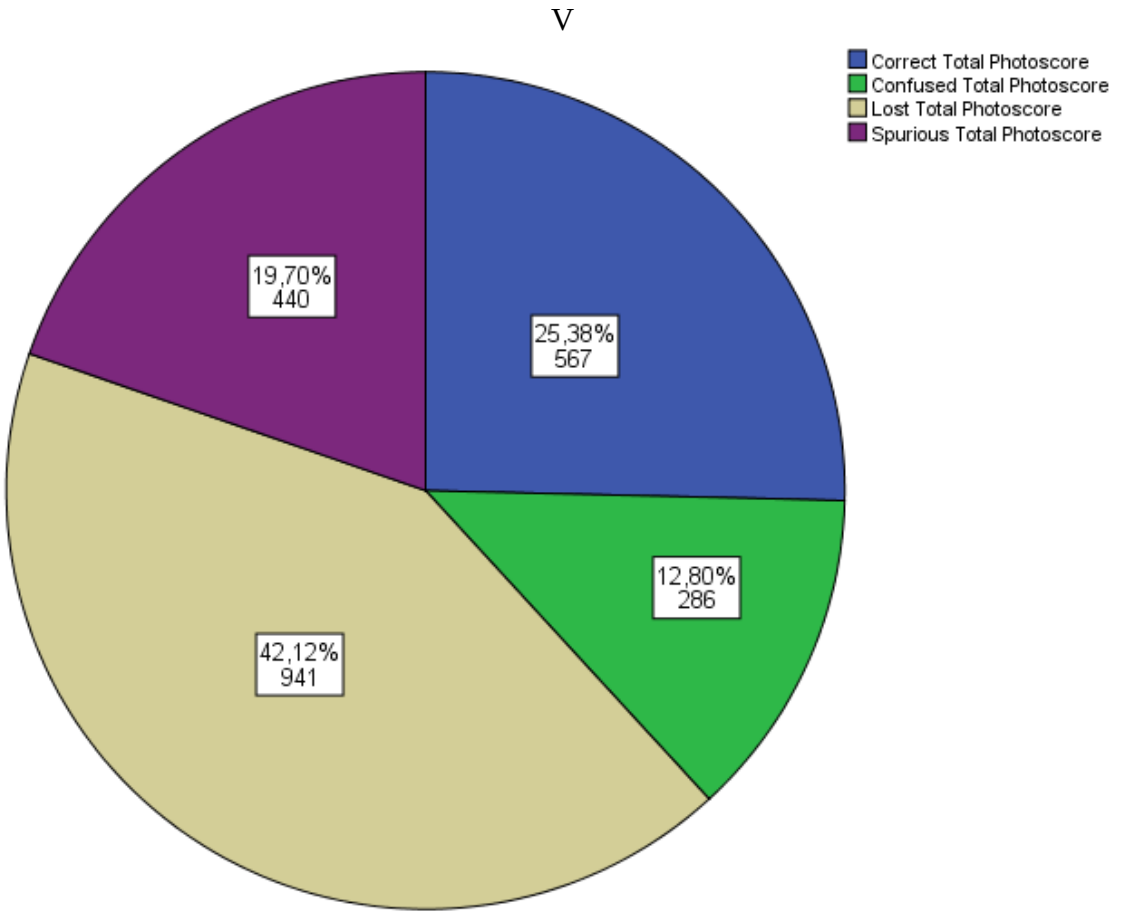


Abbildung 8: Kreisdiagramm – Typverteilung Total Photoscore

Fast die Hälfte aller Zeichen geht verloren. Dies liegt jedoch, wie gezeigt, im großen Maße an den nicht erkannten Blättern. Innerhalb der Zeichen, die der OMR-Output erzeugt, wird ungefähr die Hälfte der Zeichen korrekt erkannt, der Rest ist falsch oder wurde zu einem auch sehr großen Teil unnötigerweise erzeugt.

Insgesamt weist die deskriptive Statistik zu Photoscore auf eine für OCR/OMR-Maßstäbe schlechte Leistung hin. Dies liegt vor allem daran, dass einzelne Blätter gar nicht erkannt werden. Zieht man jedoch diese Blätter ab, bleibt das Ergebnis weiter unter den Maßstäben wie folgende Tabelle zeigt (=0% sind die Blätter die gar nicht erkannt wurde, >0% die Blätter, die zumindest als Notenblätter erkannt wurden):

Tabelle 5: Vergleich – Fehlende Notenblatterkennung Total

Bericht

Distribution Correct In Percent Total boolean Photoscore		Correct In Percent Total Photoscore	Lost In Percent Total Photoscore	Error Rate Total Photoscore
=0%	Mittelwert	,2597	98,9610	100,0000
	H	7	7	7
	Standardabweichung	,68721	2,74883	,00000
>0%	Mittelwert	55,1297	18,7971	95,3645
	H	13	13	13
	Standardabweichung	13,03600	12,29729	30,86430
Gesamtsumme	Mittelwert	35,9252	46,8545	96,9869
	H	20	20	20
	Standardabweichung	28,78305	40,45758	24,63313

Der Mittelwert bei fast überhaupt nicht erkannten Blättern liegt trivialerweise bei fast 0% für die Accuracy. Bei den anderen Blättern liegt er bei 55%. Das sind also 20% mehr als im allgemeinen Mittelwert. Dennoch ist der Wert weit unter Maßstäben und die Error Rate unabhängig davon sehr hoch. Offensichtlich werden einfach andere Fehler erzeugt (wie man oben sehen kann vor allem Spurious). Auch handelt es sich mit 7 Blättern, die eine annähernde 0%-Erkennung haben nicht um Ausreißer, sondern fast um die Hälfte. Da der Testkorpus weitestgehend zufällig, nach kategorialer Analyse, zusammengestellt wurde, kann man davon ausgehen, dass diese Blätter einen großen Anteil in der Liedblattsammlung ausmachen. Das Ignorieren dieser in der Statistik, würde zu falschen Schlüssen führen ist hier aber zur genaueren Analyse angegeben.

3.1.2.2 Noten

Da die Noten den Großteil eines Liedblattes ausmachen, zeichnet sich ein ähnliches Bild wie bei der Gesamtbetrachtung ab:

Tabelle 6: Deskriptive Statistik – Notes Photoscore

Deskriptive Statistiken

	N	Minimum	Maximum	Mittelwert	Standardab-weichung	Varianz
Total Notes	20	24,00	127,00	57,2500	30,84575	951,461
Correct In Percent Notes Photoscore	20	,00	80,77	36,8698	31,14976	970,307
Confused In Percent Notes Photoscore	20	,00	49,45	18,7032	15,67162	245,600
Lost In Percent Notes Photoscore	20	2,56	100,00	44,4270	42,00553	1764,465
Spurious In Percent Notes Photoscore	20	,00	30,43	6,6151	9,00883	81,159
Error Rate Notes Photoscore	20	19,23	100,00	69,7452	27,66313	765,249
Precision Notes Photoscore	20	,00	80,77	33,6564	28,87956	834,029
Gültige Anzahl (listenweise)	20					

So werden auch hier ca. 37% korrekt erkannt. Auch die Verteilung ist dabei ähnlich zum totalen Verhältnis.

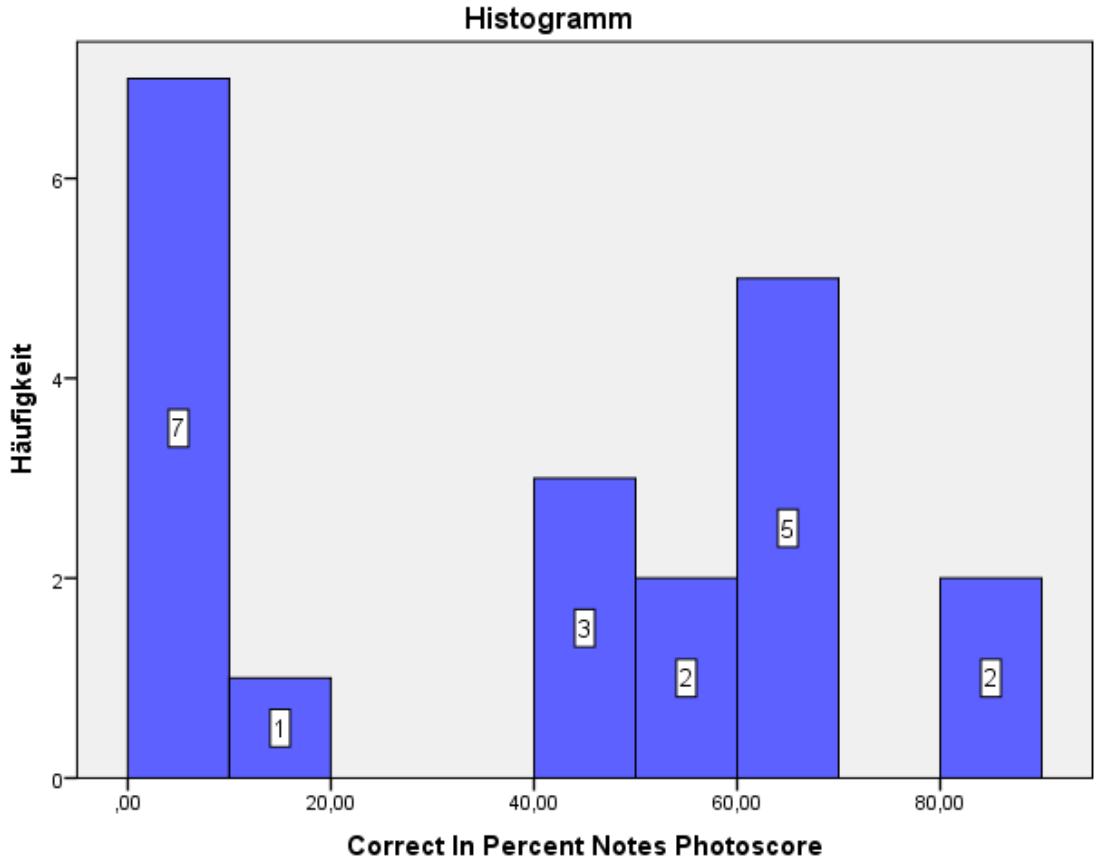


Abbildung 9: Histogramm – Correct In Percent Notes Photoscore

Jedoch lässt sich erkennen, dass die Leistung insgesamt bei den Noten etwas besser liegt als gesamt, da bei einigen Blättern eine Erkennungsrate von 80% erreicht wird. Auch hier liegt der Hauptanteil an schlechten Ergebnissen an sieben Blättern, die annähernd gar nicht erkannt werden. Dementsprechend ist die Lost-Rate, mit einem Mittelwert von 44%, sehr ähnlich zur Gesamtleistung, hoch.

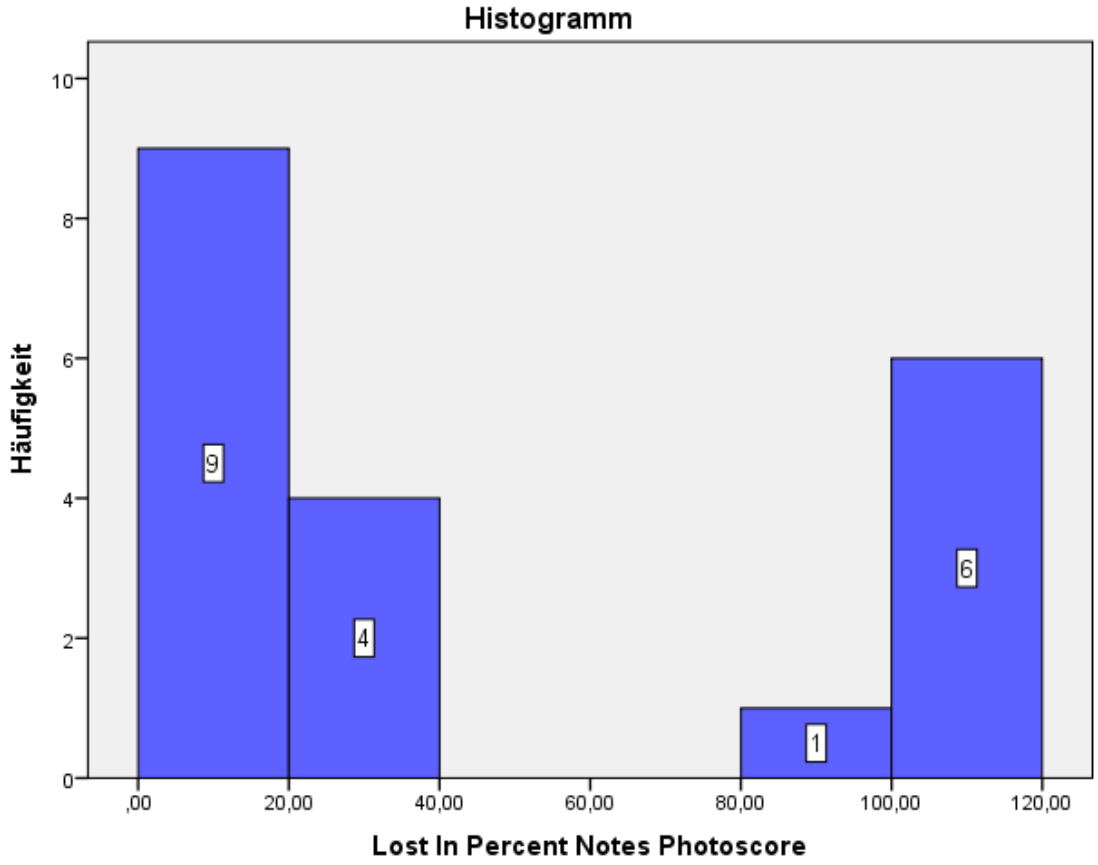


Abbildung 10: Histogramm – Lost In Percent Notes Photoscore

Auch die Confused-Rate für die Noten ist mit einem Mittelwert von 18% äquivalent zur Gesamtleistung. Die Verteilung ist lediglich gleichmäßiger und normalverteilt bis 50% nach Abzug der 0%-Raten.

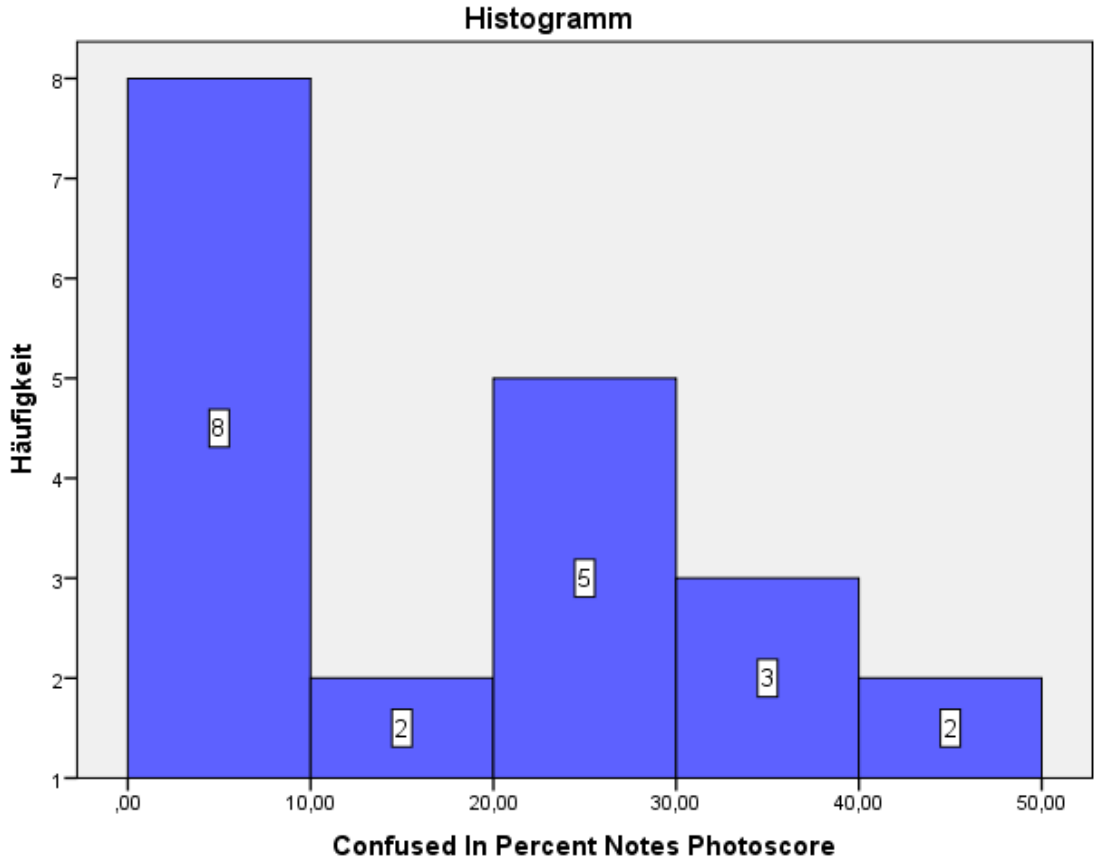


Abbildung 11: Histogramm – Confused In Percent Notes Photoscore

Ein sehr großer Unterschied liegt jedoch beim Spurios-Wert vor. Dieser liegt bei nur 6%. Im Vergleich zur Gesamtleistung von 33% machen folglich Noten einen sehr geringen Teil aus. Dies zeigt sich auch in der Verteilung:

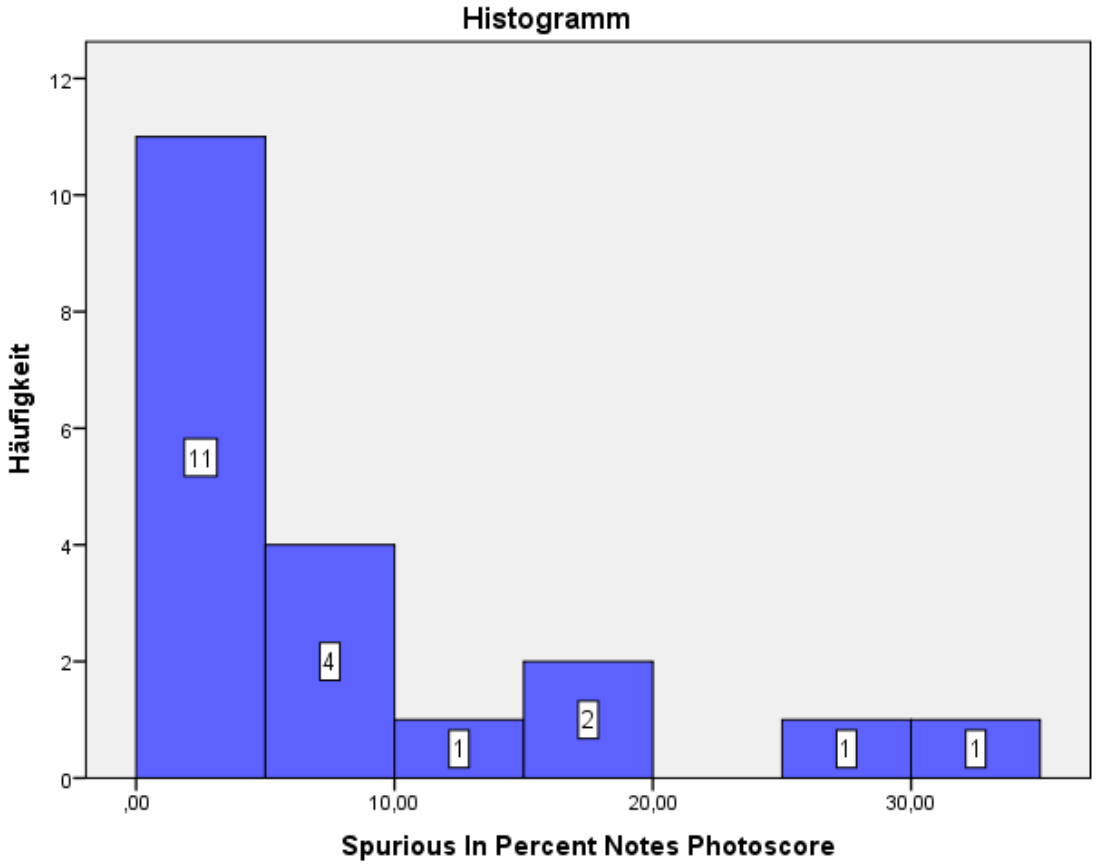


Abbildung 12: Histogramm – Spurious In Percent Notes Photoscore

Hier erkennt man, dass für die Mehrzahl der Liedblätter keine nennenswerte Erzeugung von Noise in Form von Noten vorliegt. Bei 15 Blättern unter 10%. Dies wirkt sich auch bei der Error Rate positiv aus, welche auch mit einem Mittelwert von 70% zwar weit unter Zielwerten liegt, aber für Noten auf bessere Performanz weist:

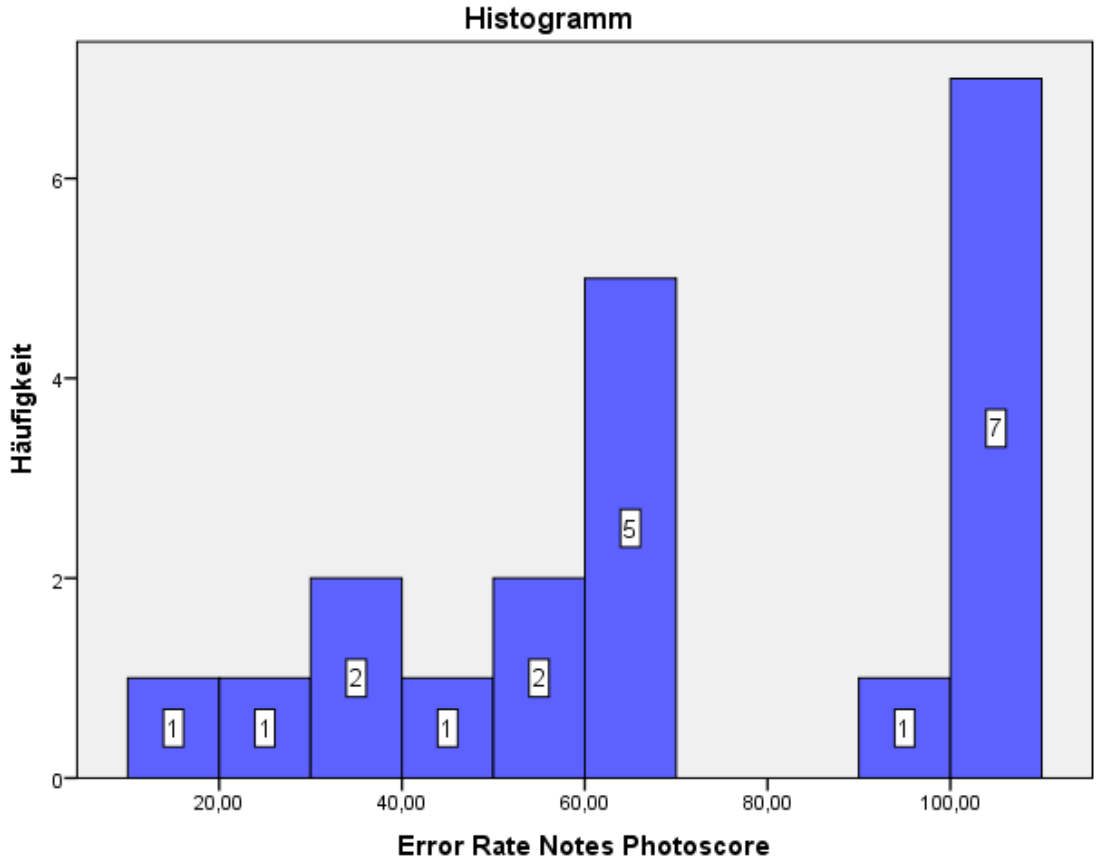


Abbildung 13: Histogramm – Error Rate Notes Photoscore

Gesamt betrachtet liegt die Verteilung der Error-Rate-Werte zwischen 60% und 120%. Bei den Noten hat man folglich eine deutlich bessere Verteilung zwischen 10% und 70%. Nur die nicht erkannten Blätter haben außerordentlich große Fehlerraten.

Die geringe Anzahl an Spurious-Fehlern hat insbesondere einen positiven Einfluss auf die Precision, mit einem Mittelwert von 34% liegt dieser 10% über der Gesamtleistung:

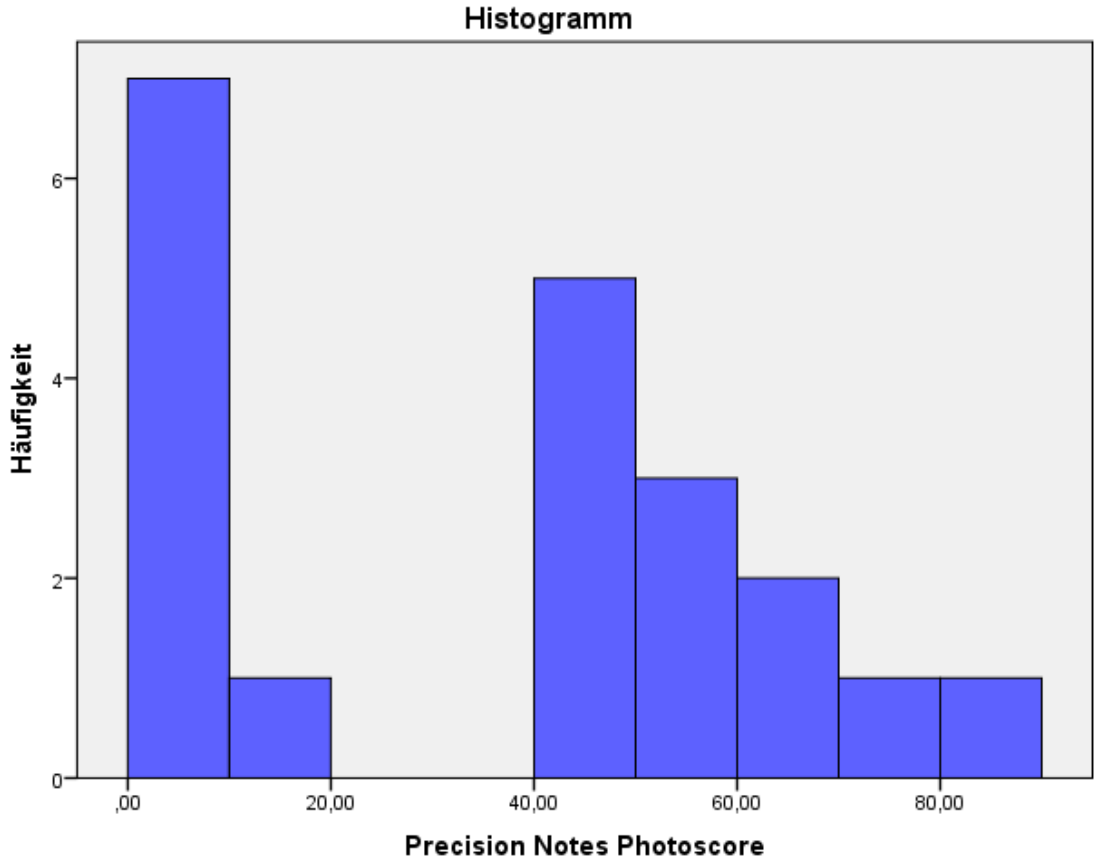


Abbildung 14: Histogramm – Precision Notes Photoscore

Die überhaupt erkannten Blätter haben dabei Werte über 40%.

Auch in der Verteilung der Erkennungskategorien macht der Spurious-Bereich den größten Unterschied zur Gesamtleistung. Für die restlichen Werte ist das Ergebnis annähernd äquivalent:

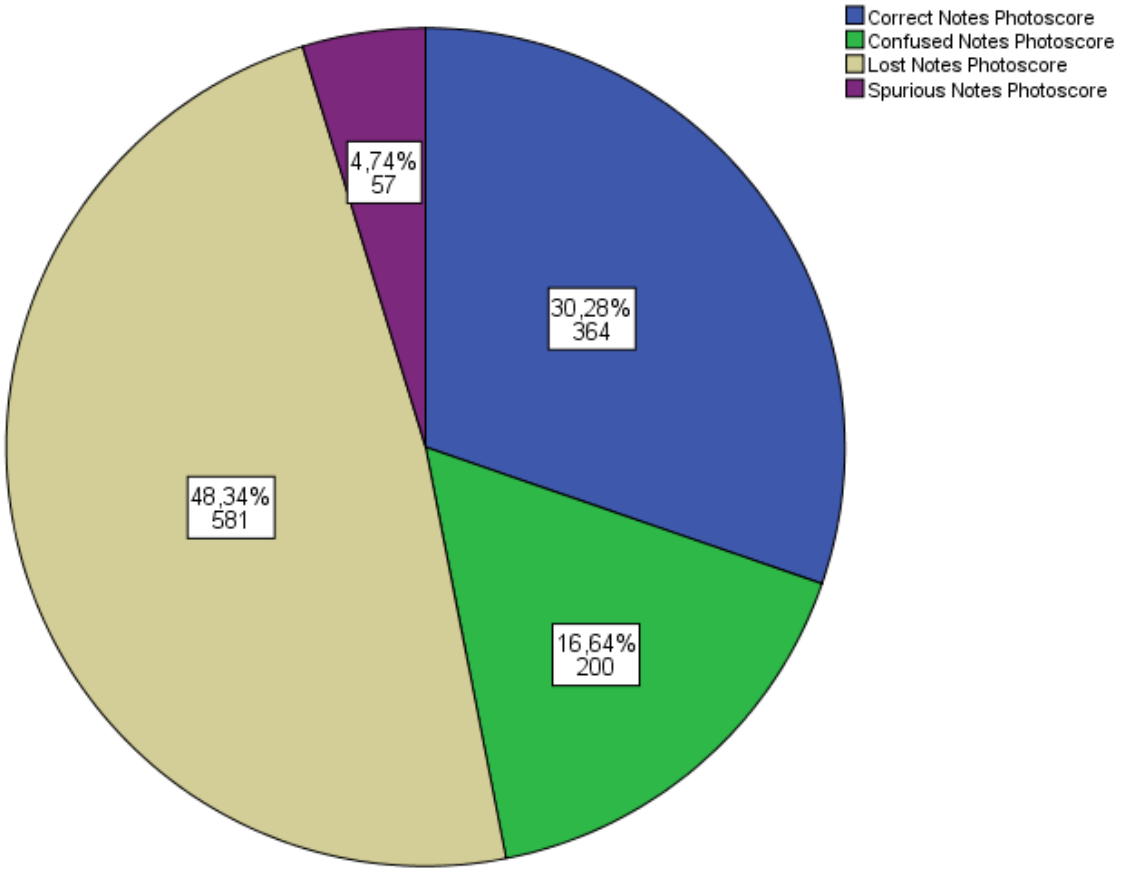


Abbildung 15: Kreisdiagramm – Typverteilung Notes Photoscore

Auch hier wird deutlich, dass das Hauptproblem bei den nicht erkannten Blättern liegt und den daraus folgenden großen Lost-Anteilen, hier etwa die Hälfte aller Zeichen. Die Spurious-Zeichen im OMR-Output sind sehr gering im Bereich der Noten. Jedoch wird ein Drittel der überhaupt erkannten Zeichen falsch erkannt.

Auch in der Kategorie der Noten kann man die gar nicht erkannten Blätter ignorieren und die wichtigsten Metriken analysieren:

Tabelle 7: Vergleich – Fehlende Notenblatterkennung Notes

Bericht

Distribution Correct In Percent Total boolean Photoscore		Correct In Percent Notes Photoscore	Lost In Percent Notes Photoscore	Error Rate Notes Photoscore
=0%	Mittelwert	,0000	98,8722	100,0000
	H	7	7	7
	Standardabweichung	,00000	2,98393	,00000
>0%	Mittelwert	56,7228	15,1104	53,4542
	H	13	13	13
	Standardabweichung	17,78647	11,35934	19,75283
Gesamtsumme	Mittelwert	36,8698	44,4270	69,7452
	H	20	20	20
	Standardabweichung	31,14976	42,00553	27,66313

Der Mittelwert für die Accuracy liegt dann zwingenderweise, ähnlich zur Betrachtung aller Zeichen gesamt, 20% über dem allgemeinen Wert. Der Lost-Wert ist trivialerweise geringer als gesamt. Allerdings liegt er mit 15% doch noch recht hoch, also 15% aller Noten gehen auch verloren, selbst wenn das Blatt und die Notenzeilen grundsätzlich erkannt wurden. Auch die Fehlerrate liegt bei 53% obschon die Spurious-Werte sehr gering sind, d.h. wenn Photoscore ein Liedblatt grundsätzlich erkennt, gehen immer noch fast die Hälfte der Zeichen verloren oder sind falsch.

Insgesamt kann man festhalten, dass die Leistung in Bezug auf die Noten geringfügig besser ist, aber noch immer weit unter angestrebten Zielwerten für alle relevanten Maße.

3.1.2.3 Noten mit Sonderzeichen

Eine untersuchte Sonderkategorie an musikalischen Zeichen sind Noten mit Sonderzeichen, also mit Vorzeichen, punktiert oder Ähnlichem. Insgesamt gibt es im gesamten Korpus 90 dieser Zeichen, was ca. 5 % entspricht. Auch beinhalten nur 13 Blätter diese Zeichen überhaupt. Die Bedeutung ist demnach eher gering. Die deskriptive Statistik zu diesen lässt sich in der folgenden Tabelle zusammenfassen:

Tabelle 8: Deskriptive Statistik – Notes With Special Characters

Deskriptive Statistiken

	N	Minimum	Maximum	Mittelwert	Standardabweichung	Varianz
Total Notes With Special Characters	20	,00	21,00	4,5000	5,72621	32,789
Correct In Percent Notes With Special Characters Photoscore	13	,00	100,00	33,6172	36,65254	1343,409
Confused In Percent Notes With Special Characters Photoscore	13	,00	100,00	19,5849	28,86928	833,435
Lost In Percent Notes With Special Characters Photoscore	13	,00	100,00	46,7979	45,48973	2069,316
Spurious In Percent Notes With Special Characters Photoscore	13	,00	1100,00	145,9554	306,06477	93675,646
Error Rate Notes With Special Characters Photoscore	13	44,44	1200,00	212,3382	317,97744	101109,655
Precision Notes With Special Characters Photoscore	13	,00	60,00	21,8116	23,43979	549,424
Gültige Anzahl (listenweise)	13					

Ein Liedblatt besitzt im Durchschnitt ca. 5 dieser Zeichen. Sowohl die Erkennungsrate als auch die Lost- und Confused-Rate verhalten sich äquivalent wie bei den Noten, als auch Gesamt.

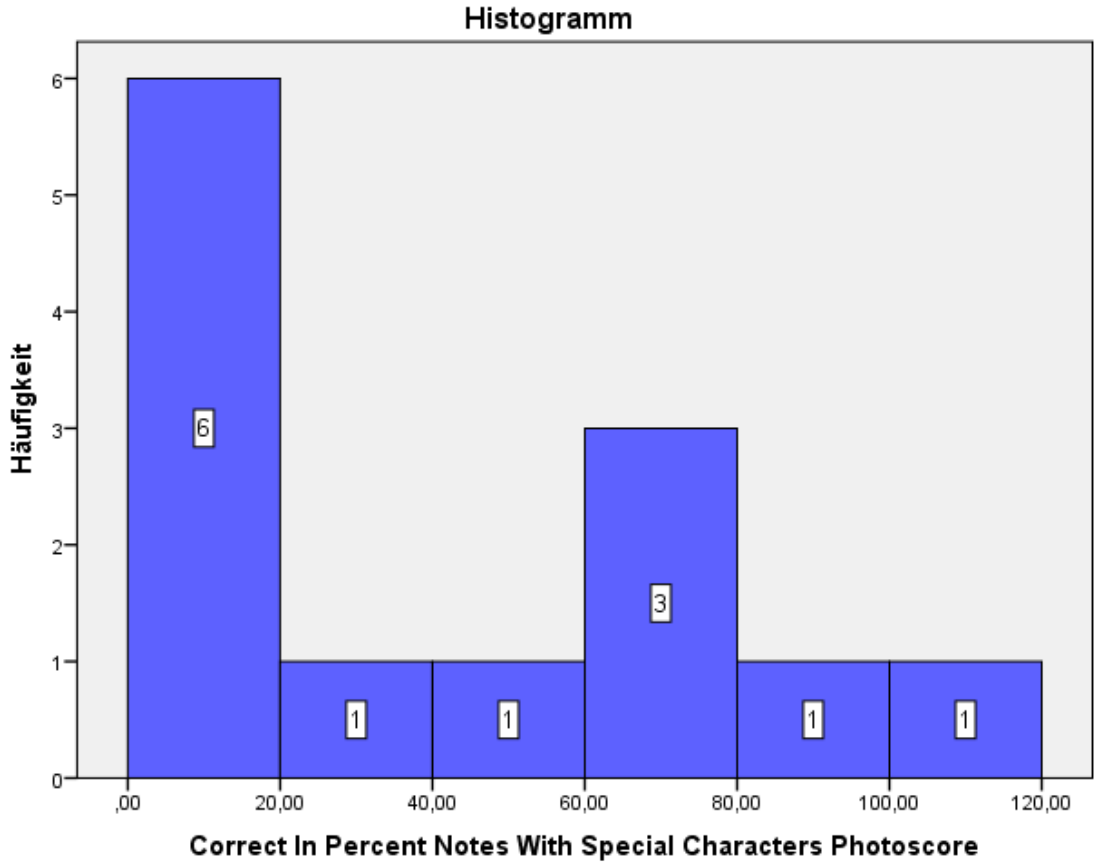


Abbildung 16: Histogramm – Correct In Percent Notes With Special Characters Photoscore

Aufgrund der geringen Anzahl an diesen Zeichen ist die Performanz jedoch geringfügig besser im oberen Bereich, will heißen es gibt einige Blätter die fast zu 100% erkannt werden für diese Zeichentypen.

Ein großer Unterschied lässt sich jedoch bei den Spurious-Werten feststellen. Mit einem Mittelwert von 146% ist diese Rate sehr hoch. Aufgrund der geringen Gesamtanzahl an Noten mit Sonderzeichen ist dieser Wert jedoch absolut gesehen weniger negativ. Im Schnitt werden einem Notenblatt also 1.5-mal so viele Zeichen dieser Sorte hinzugefügt wie tatsächlich vorliegen. Man muss beachten, dass wie in der Methodologie schon erklärt wurde, für jede Kategorie das Blatt unabhängig von anderen Zeichen betrachtet wird. Ob also andere Zeichentypen als Noten mit Sonderzeichen verwechselt wurden oder tatsächliche Noise erzeugt wird ist hierbei nicht ersichtlich. Bei genauerer Betrachtung lässt sich jedoch feststellen, dass ein großer Teil dieses Ergebnis durch zwei

Ausreißer entsteht, welche Raten von 400 bzw. 1000% haben. Eliminiert man diese Ausreißer in einem Histogramm mit den absoluten Zahlen, kann man die Verteilung genauer betrachten:

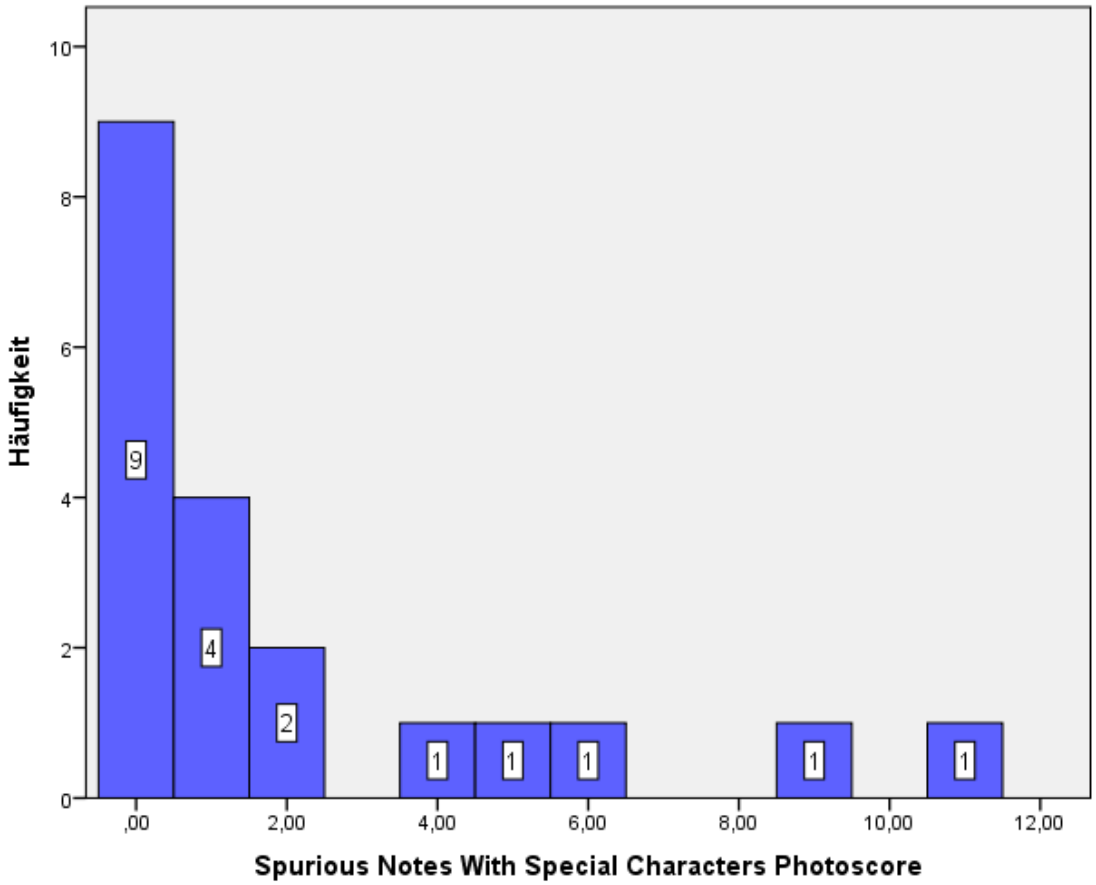


Abbildung 17: Histogramm – Spurious Notes With Special Characters Photoscore

Die absoluten Zahlen stellen das Verhältnis wieder klar dar. So werden zu der Mehrzahl der Blätter nur geringfügig viele Zeichen als Noise erzeugt. Tatsächlich reicht aber schon eine geringe Zahl um prozentual betrachtet sehr große Spurious-Raten zu erzeugen.

Insgesamt ist dieser Zeichentyp zu vernachlässigen und weist nur bezüglich der Spurious-Rate eine Abweichung zu den Noten auf. Dies zeigt auch das Kreisdiagramm über den gesamten Korpus:

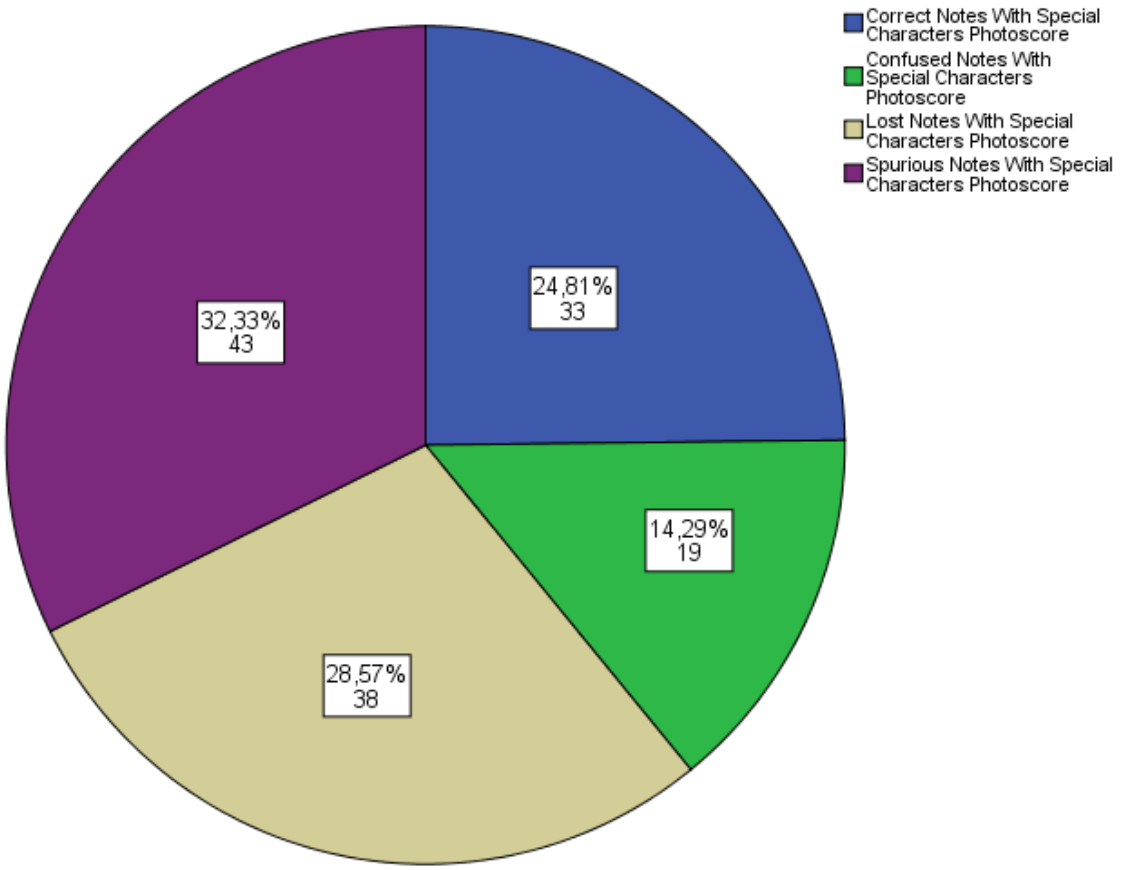


Abbildung 18: Kreisdiagramm – Typverteilung Notes With Special Characters Photoscore

Bis auf die Spurious-Werte, welche bei gar nicht erkannten Blättern natürlich 0% ergeben, ist die Performanz bei den grundsätzlich erkannten Blättern geringfügig besser:

Tabelle 9: Vergleich – Fehlende Notenblatterkennung Notes With Special Characters

Bericht

Distribution General Recognition boolean Photoscore		Correct In Percent Notes With Special Characters Photoscore	Lost In Percent Notes With Special Characters Photoscore	Error Rate Notes With Special Characters Photoscore
=0%	Mittelwert	,0000	100,0000	100,0000
	H	4	4	4
	Standardabweichung	,00000	,00000	,00000
>0%	Mittelwert	48,5582	23,1526	262,2663
	H	9	9	9
	Standardabweichung	34,62539	32,55383	377,55811
Gesamtsumme	Mittelwert	33,6172	46,7979	212,3382
	H	13	13	13
	Standardabweichung	36,65254	45,48973	317,97744

3.1.2.4 Pausen

Mit einem Anteil von 2,3 % treten Pausen sehr selten im Korpus auf. Es gibt nur 14 Blätter die überhaupt Pausen enthalten. Tatsächlich erkennt man an der deskriptiven Statistik das ähnliche Phänomen wie bei der obigen Kategorie, nämlich sehr große Spurious-Raten:

Tabelle 10: Deskriptive Statistik – Rests Photoscore

Deskriptive Statistiken

	N	Minimum	Maximum	Mittelwert	Standardab-weichung	Varianz
Total Rests	20	,00	7,00	2,0500	2,08945	4,366
Correct In Percent Rests Photoscore	14	,00	100,00	27,0918	41,28097	1704,118
Confused In Percent Rests Photoscore	14	,00	100,00	25,9184	38,31255	1467,852
Lost In Percent Rests Photoscore	14	,00	100,00	46,9898	46,53300	2165,320
Spurious In Percent Rests Photoscore	14	,00	900,00	321,6224	325,99457	106272,457
Error Rate Rests Photo-score	14	100,00	1000,00	393,6735	316,53588	100194,965
Precision Rests Photo-score	14	,00	44,44	6,1634	12,23187	149,619
Gültige Anzahl (listen-weise)	14					

Neben der sehr hohen Spurious-Rate mit einem Mittelwert von 321% ist aber auch die Erkennungsrate mit 27% sehr schlecht. Auch hier wirken sich jedoch geringfügige Fehlleistungen stark aus, da ein Blatt im Schnitt 2 Pausen hat, reicht es wenn nur eine nicht erkannt wird um eine Accuracy von 50% auszugeben. Gleichzeitig ist es „leichter“ 100% zu erreichen:

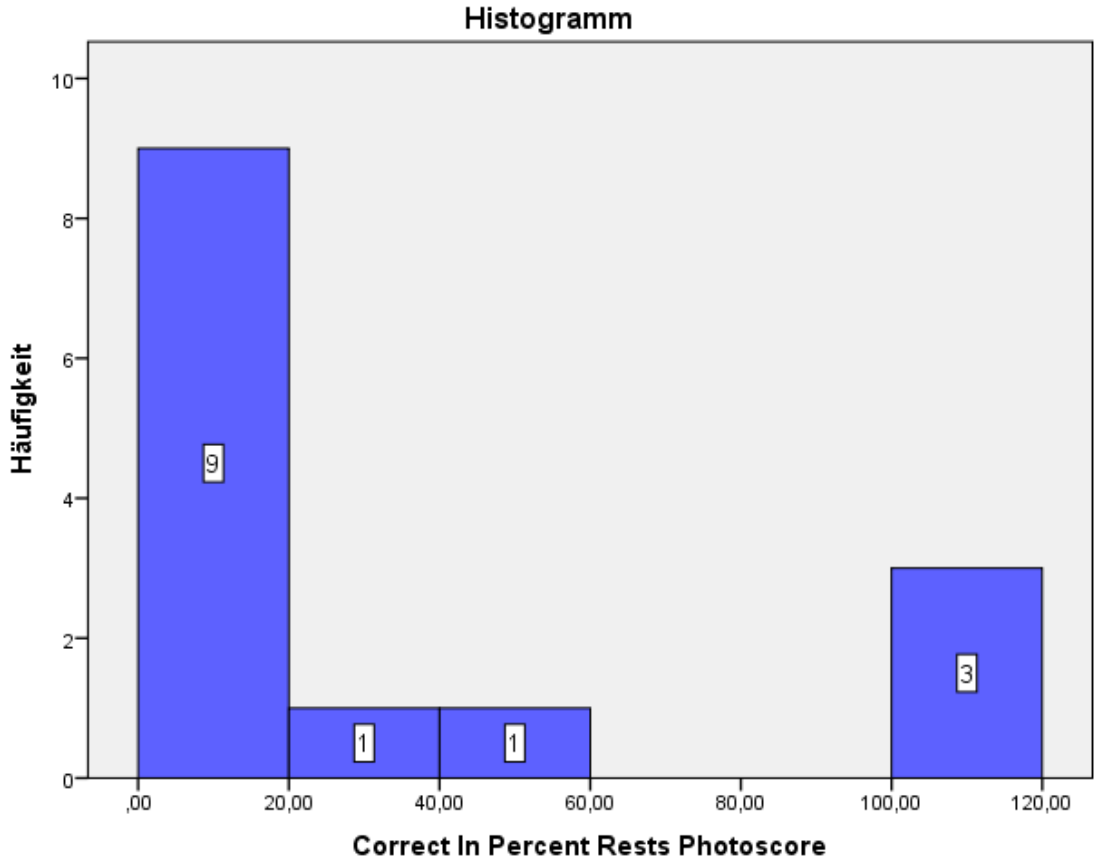


Abbildung 19: Histogramm – Correct In Percent Rests Photoscore

Wie der Mittelwert vermuten lässt ist die Verteilung von Spurious-Werten sehr schlecht:

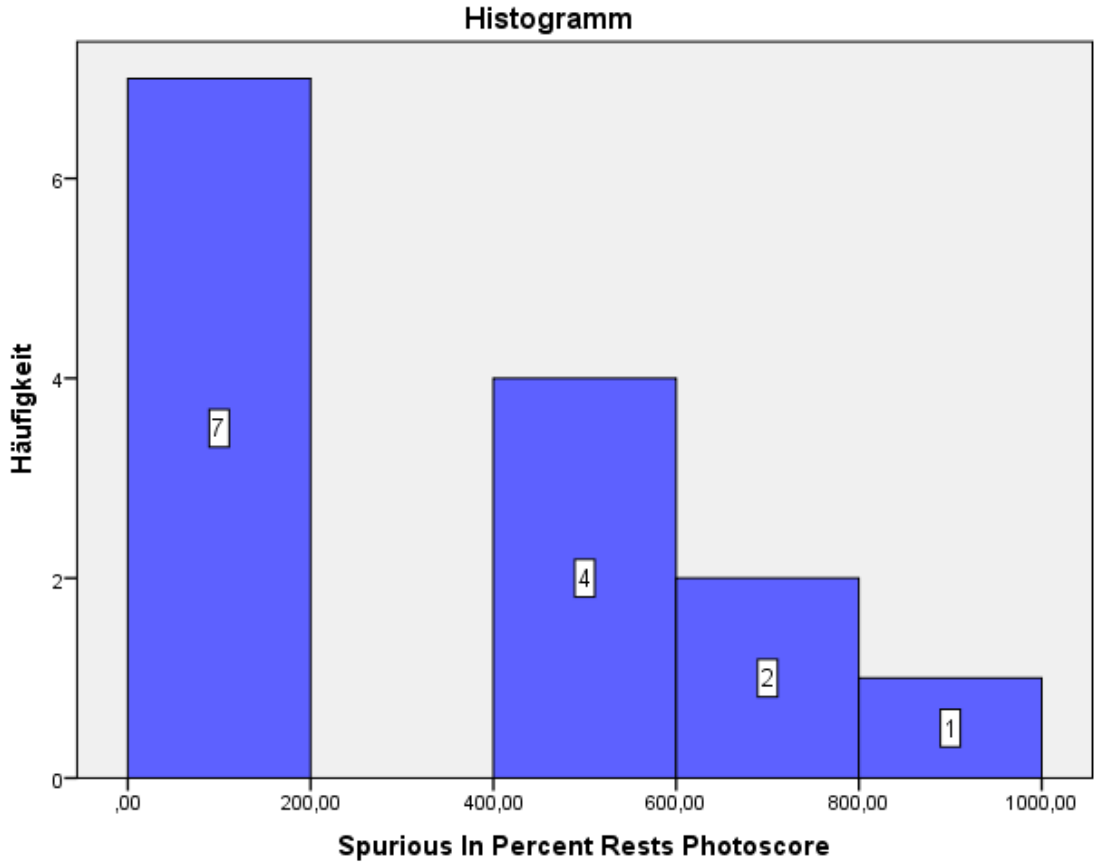


Abbildung 20: Histogramm – Spurious In Percent Rests Photoscore

Ungefähr die Hälfte aller Blätter, die Pausen enthalten erzeugen folglich mit Photoscore einen Noise-Anteil von über 400%. Die absoluten Zahlen relativieren dabei die Verhältnisse:

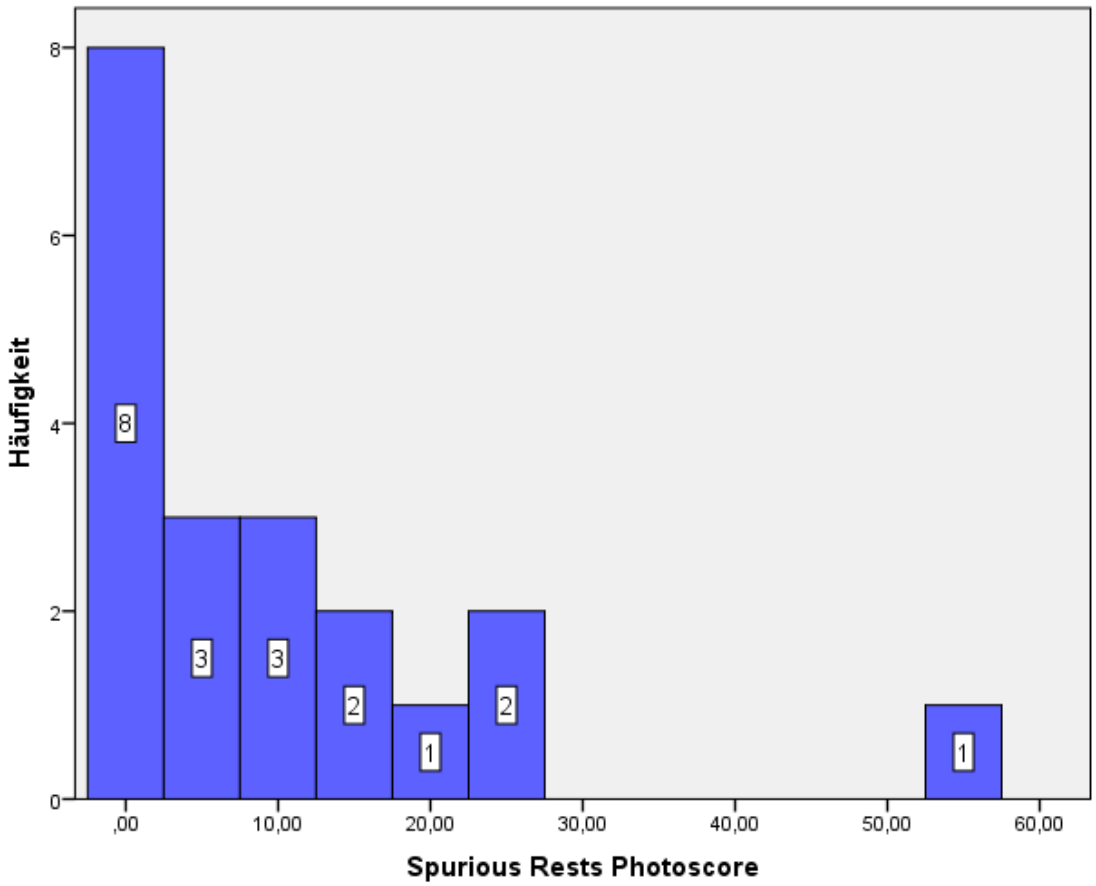


Abbildung 21: Histogramm – Spurious Rests Photoscore

Die meisten Blätter fügen tatsächlich keine neuen Pausen ein (bzw. verwechseln andere Zeichentypen als Pausen). Jedoch handelt es sich dabei wieder um die Blätter die von Photoscore gar nicht erkannt wurden und somit gar keine Noise erzeugen. Die meisten Blätter fügen im Schnitt 5 – 25 Pausen hinzu.

Im Kreisdiagramm für die Fehlertypen dominieren die Spurious-Zeichen:

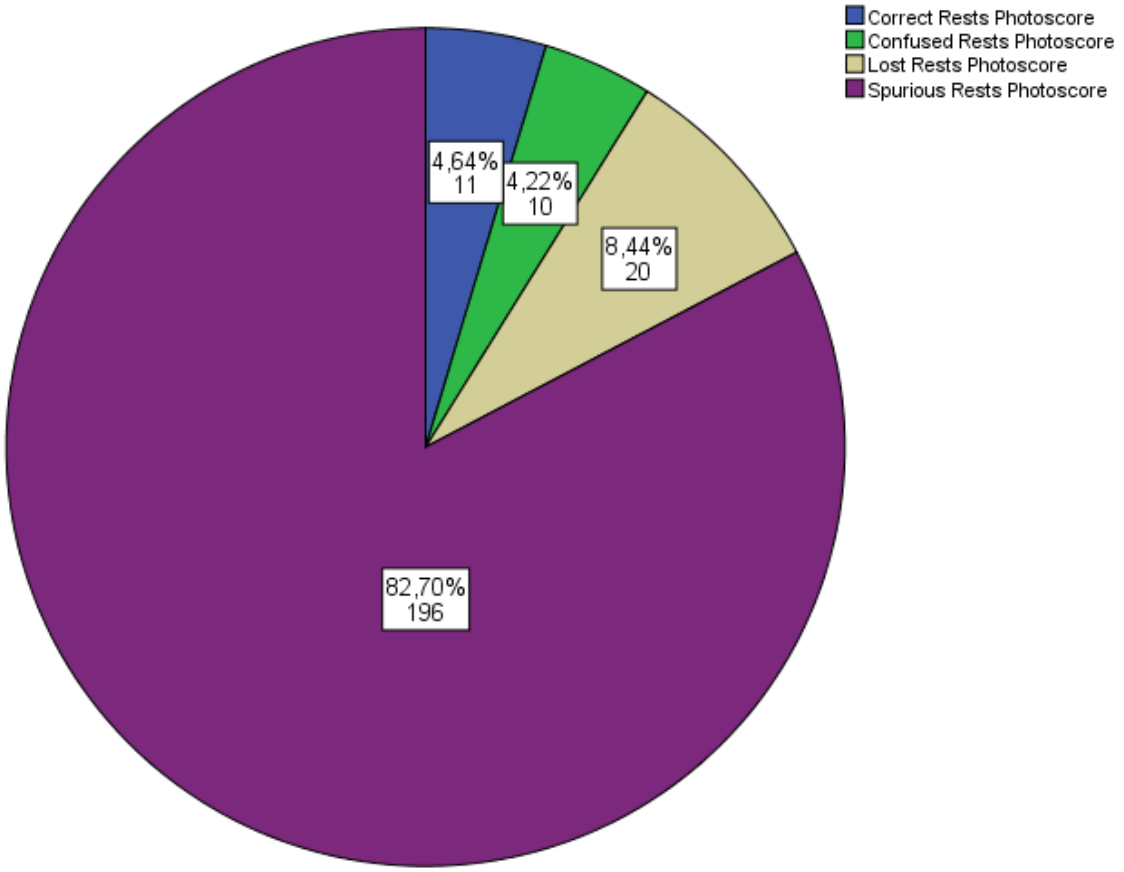


Abbildung 22: Kreisdiagramm – Typverteilung Rests Photoscore

Eine extrem große Anzahl von Zeichen wird offensichtlich als Pausen interpretiert. Auch muss man beachten, dass Photoscore leere oder unvollständige Zeilen automatisch mit Pausen befüllt. Dies erklärt die hohen Werte. Nichtsdestotrotz müssen diese Pausen bei einer tatsächlichen Auszeichnung entfernt werden. Auch hier wird noch das Performanz-Ergebnis im Vergleich der gar nicht erkannten Blätter zu den grundsätzlich erkannten Blättern angegeben:

Tabelle 11: Vergleich – Fehlende Notenblatterkennung Rests

Bericht

Distribution General Recognition boolean Photoscore		Correct In Percent Rests Photoscore	Lost In Percent Rests Photoscore	Error Rate Rests Photoscore
=0%	Mittelwert	,0000	100,0000	110,0000
	H	5	5	5
	Standardabweichung	,00000	,00000	22,36068
>0%	Mittelwert	42,1429	17,5397	551,2698
	H	9	9	9
	Standardabweichung	45,33886	28,04677	290,39882
Gesamtsumme	Mittelwert	27,0918	46,9898	393,6735
	H	14	14	14
	Standardabweichung	41,28097	46,53300	316,53588

Die Erkennungsrate ist jedoch mit 42% auch bei diesen Blättern, isoliert betrachtet, unterdurchschnittlich. Die Verlustrate mit 17% auch noch hoch. Die große Error Rate wird hauptsächlich über die Spurious-Werte bestimmt.

3.1.2.5 Taktstriche

Mit 16% ist der Anteil der Taktstriche am Gesamt-Pool von Zeichen am zweitgrößten nach den Noten. Im Gegensatz zu den anderen Sonder-Kategorien kommen diese Zeichen auch in jedem der 20 Liedblätter vor. Zuerst eine Zusammenfassung der deskriptiven Statistik:

Tabelle 12: Deskriptive Statistik – Bar Lines Photoscore

Deskriptive Statistiken

	N	Minimum	Maximum	Mittelwert	Standardab-weichung	Varianz
Total Bar Lines	20	1,00	34,00	14,4000	9,82692	96,568
Correct In Percent Bar Line Photoscore	20	,00	90,48	36,6561	37,65803	1418,127
Confused In Percent Bar Line Photoscore	20	,00	100,00	8,9785	22,92948	525,761
Lost In Percent Bar Line Photoscore	20	,00	100,00	54,3654	42,76162	1828,556
Spurious In Percent Bar Line Photoscore	20	,00	600,00	118,7552	152,76663	23337,643
Error Rate Bar Line Photo-score	20	44,44	700,00	182,0991	142,67826	20357,086
Precision Bar Line Photo-score	20	,00	63,64	15,0082	17,05857	290,995
Gültige Anzahl (listen-weise)	20					

Ein Notenblatt hat im Schnitt 14 Taktstriche. Auch hier ist die Erkennungsrate äquivalent zu den bisherigen Kategorien bei ca. 37%. Dies wird beeinflusst durch große Lost-Raten von 100% bei den Sieben Blättern, die gar nicht von Photoscore als Notenblätter erkannt wurden.

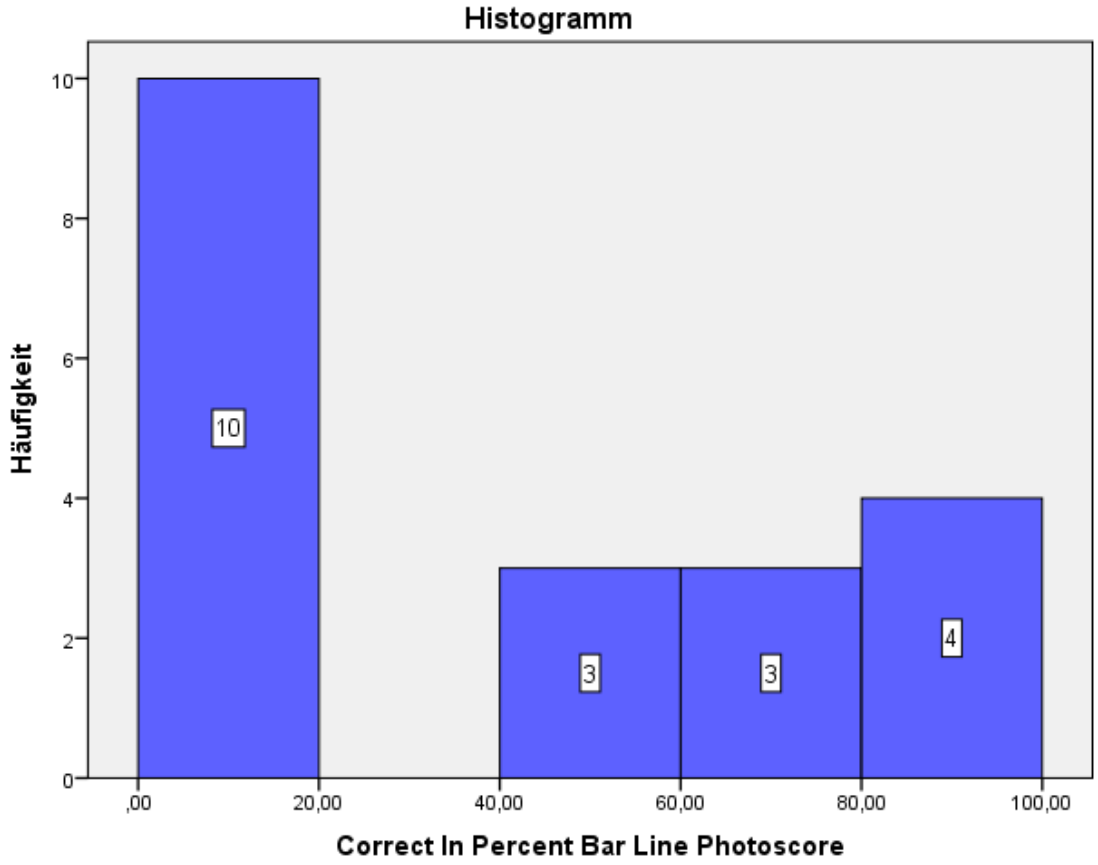


Abbildung 23: Histogramm – Correct In Percent Bar Lines Photoscore

Auch hier gibt es jedoch im Gegensatz zu den Noten auch Blätter die über 80% Erkennungsraten haben.

Wie bei jeder Sonderkategorie liegt der größte Unterschied zu den Noten im Bereich der Spurious-Werte vor, mit einem sehr großen Mittelwert von 118%. Da im Schnitt jedoch ein Liedblatt 14 Taktstriche besitzt, kann man nicht von einer ähnlichen Verzerrung wie bei den Pausen oder den Noten mit Sonderzeichen sprechen. Es handelt sich um relevante Noise:

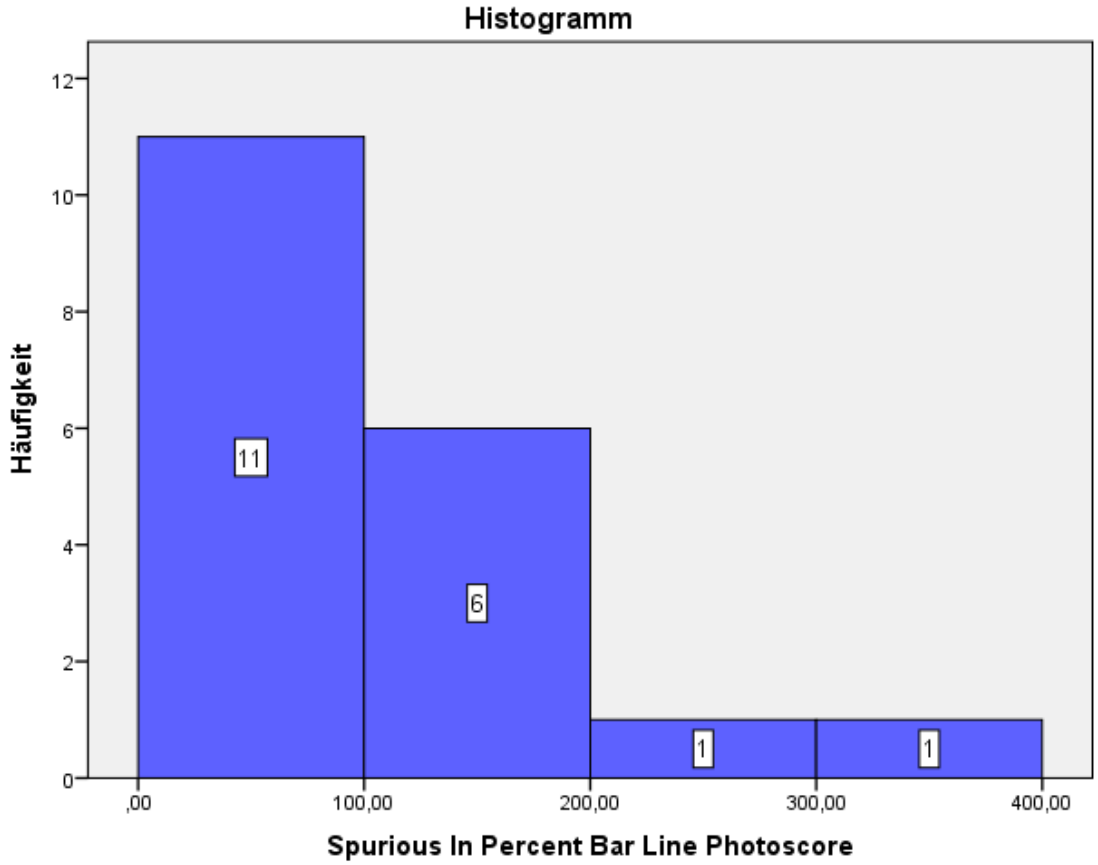


Abbildung 24: Histogramm – Spurious In Percent Bar Lines Photoscore

Ein Ausreißer ist im Histogramm beim Maximum von 600% nicht vermerkt. Auch hier liefert das Histogramm für die absoluten Werte eine weniger verzerrte Einsicht:

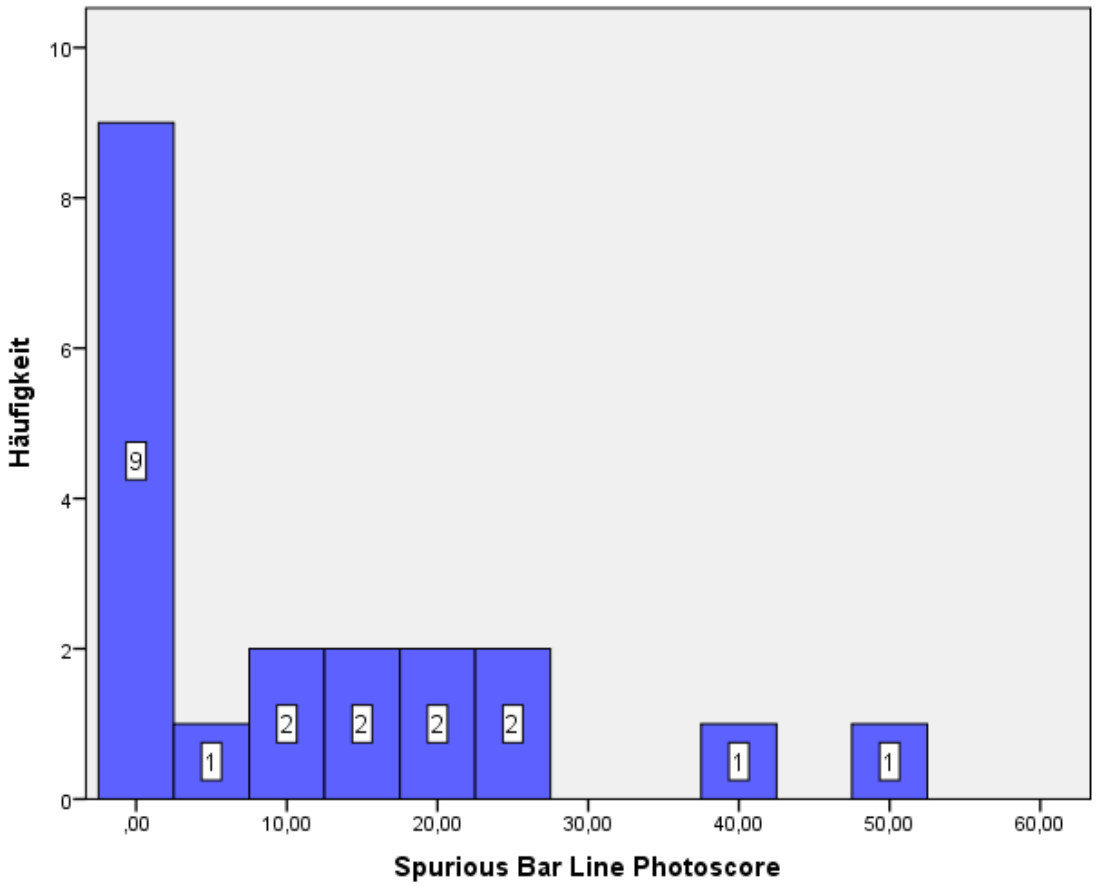


Abbildung 25: Histogramm – Spurious Bar Lines Photoscore

Fast die Hälfte der Liedblätter produziert genauso viele Taktstriche wie es eigentlich enthält noch als überflüssige Noise. Dies entspricht hauptsächlich zwischen 5 und 25 Zeichen. Der Wert lässt sich ähnlich wie zu den Pausen unter anderem dadurch erklären, dass Photoscore bei leeren oder unvollständigen Zeilen eigene Taktstriche einfügt. Auch werden offensichtlich viele Zeichen anderer Kategorien als Taktstriche interpretiert.

Im Kreisdiagramm der OMR-Typen dominieren folglich die Spurious-Zeichen:

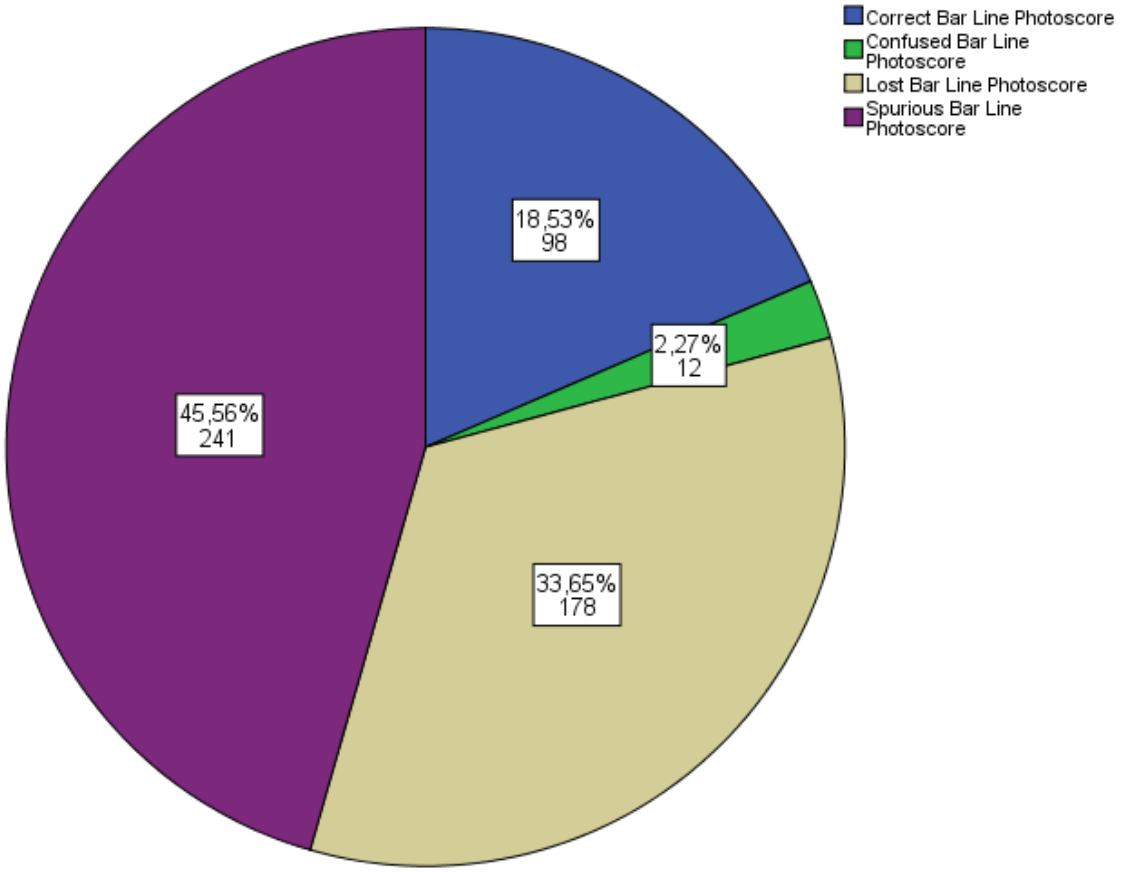


Abbildung 26: Kreisdiagramm – Typverteilung Bar Lines Photoscore

Taktstriche werden selten falsch erkannt, jedoch sind nur 20% bei der OMR-Erkennung korrekt. Der Rest ist verloren oder Noise. Die Leistung ist demnach kaum brauchbar.

Sortiert man die Blätter raus, die Photoscore nicht als Notenblätter erkannt hat, ist das Ergebnis nur geringfügig anders:

Tabelle 13: Vergleich – Fehlende Notenblatterkennung Bar Lines Photoscore

Bericht

Distribution General Recognition boolean Photoscore		Correct In Percent Bar Line Photoscore	Lost In Percent Bar Line Photoscore	Error Rate Bar Line Photoscore
=0%	Mittelwert	,0000	100,0000	100,0000
	H	7	7	7
	Standardabweichung	,00000	,00000	,00000
>0%	Mittelwert	56,3940	29,7930	226,3063
	H	13	13	13
	Standardabweichung	32,24148	32,03577	161,81192
Gesamtsumme	Mittelwert	36,6561	54,3654	182,0991
	H	20	20	20
	Standardabweichung	37,65803	42,76162	142,67826

Die Erkennungsrate steigert sich zwar äquivalent zu den anderen Kategorien um 20%. Lost-Rate und Error-Rate sind aber gleichbleibend schlecht, im Falle der Error Rate sogar schlechter, da die großen Spurious-Werte stärker gewichtet sind.

3.1.2.6 Sonderzeichen

Unter Sonderzeichen versteht man Violin- oder Basschlüssel und Tonartangaben. 13% aller Zeichen sind dieser Art. Jedes Liedblatt enthält eine gewisse Anzahl dieser. Folgende Tabelle fasst die Leistung von Photoscore in Bezug auf diese Zeichen zusammen:

Tabelle 14: Deskriptive Statistik – Special Characters Photoscore

Deskriptive Statistiken

	N	Minimum	Maximum	Mittelwert	Standardabweichung	Varianz
Total Special Characters	20	3,00	30,00	11,5000	7,61923	58,053
Correct In Percent Special Characters Photoscore	20	,00	70,00	26,7858	25,17908	633,986
Confused In Percent SpecialCharacters Photoscore	20	,00	88,24	22,7391	28,01144	784,641
Lost In Percent Special Characters Photoscore	20	,00	100,00	50,4751	37,98417	1442,797
Spurious In Percent Special Characters Photoscore	20	,00	100,00	13,8571	27,55368	759,205
Error Rate Special Characters Photoscore	20	32,00	166,67	87,0714	34,22137	1171,102
Precision Special Characters Photoscore	20	,00	68,00	23,7400	23,07546	532,477
Gültige Anzahl (listenweise)	20					

Ein Liedblatt besitzt im Schnitt 12 dieser Zeichen und mindestens drei. Die Erkennungsrate ist jedoch mit 26% um ca. 10% schlechter als bei den meisten anderen Kategorien. Die Erkennung dieser Zeichen ist also noch schwächer.

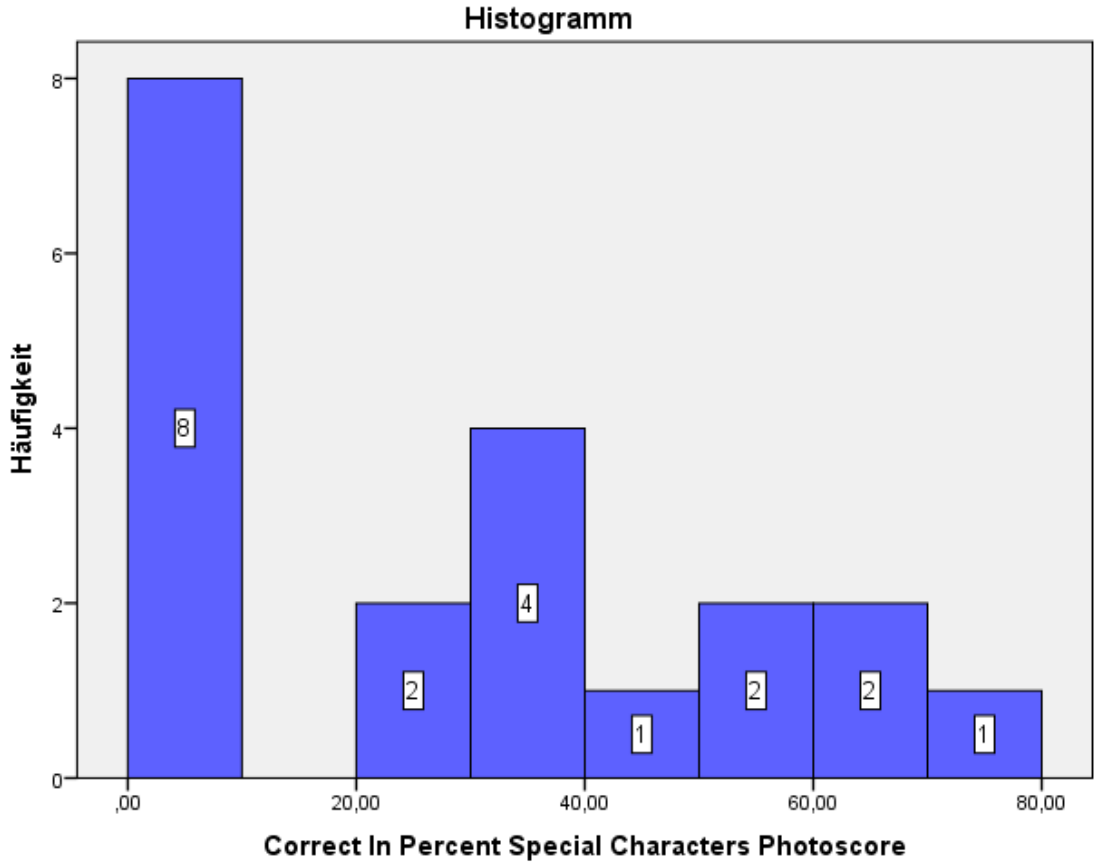


Abbildung 27: Histogramm – Correct In Percent Special Characters Photoscore

Auch hier wiegen die gar nicht erkannten Notenblätter sich negativ auf die Erkennungsrate aus. Die restlichen Blätter streuen zwischen 20% und 80%.

Im Gegensatz zu den anderen Sonder-Kategorien sind die Spurious-Werte nicht außerordentlich größer und auch die anderen Fehlertypen wie die Confused- und Lost-Raten sind ähnlich zu den Gesamt-Wertungen. Der Spurious-Anteil ist mit einem Mittelwert von 13% sogar fast 20% geringer als für alle Kategorien zusammen. Diese Verteilung zeigt sich auch im Kreisdiagramm für Typen:

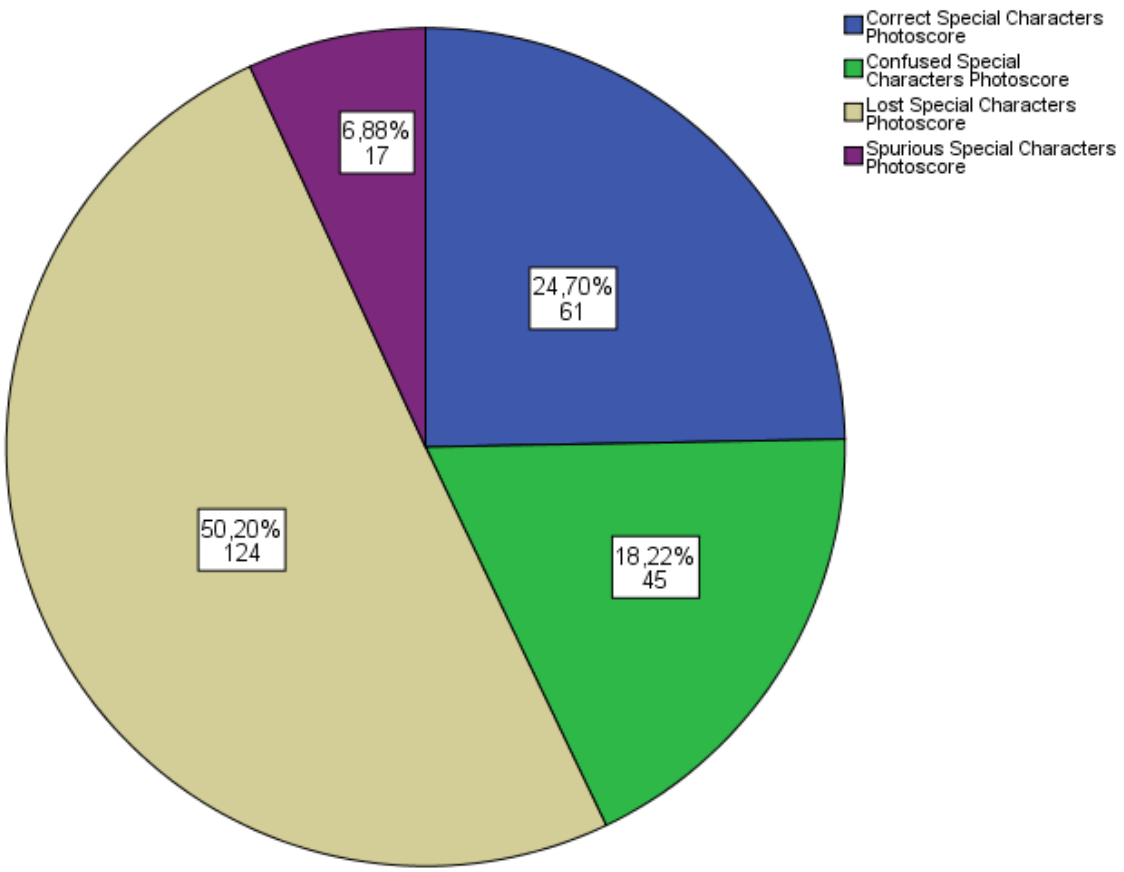


Abbildung 28: Kreisdiagramm – Typverteilung Special Characters Photoscore

Die Grafik zeigt eine große Ähnlichkeit zu der, der Noten. Für Sonderzeichen gilt also, dass der Großteil verloren geht, die Erkennungsrate etwas schlechter ist als in den anderen Kategorien, jedoch sehr viel weniger Noise erzeugt wird. Auch hier liegen hohe Lost-Werte vor. Alle Performanz-Parameter verbessern sich aber auch hier wenn man die nicht erkannten Blätter aussortiert:

Tabelle 15: Vergleich – Fehlende Notenblatterkennung Special Characters

Bericht

Distribution General Recognition boolean Photoscore		Correct In Percent Special Characters Photoscore	Lost In Percent Special Characters Photoscore	Error Rate Special Characters Photoscore
=0%	Mittelwert	4,7619	95,2381	95,2381
	H	7	7	7
	Standardabweichung	12,59882	12,59882	12,59882
>0%	Mittelwert	38,6448	26,3719	82,6739
	H	13	13	13
	Standardabweichung	22,11659	20,17041	41,41286
Gesamtsumme	Mittelwert	26,7858	50,4751	87,0714
	H	20	20	20
	Standardabweichung	25,17908	37,98417	34,22137

Auch hier ist bei den Blättern, die zumindestens als Notenblätter erkannt wurden der Mittelwert für Accuracy und Lost In Percent besser, aber noch weit unter Zielwert. Die Error Rate ist gleichbleibend.

3.1.2.7 Kategoriale Vergleiche

In den vorherigen Abschnitten wurde grob auf Unterschiede und Besonderheiten der einzelnen Zeichentypen eingegangen. Hier sollen diese über deskriptive Mittelwertvergleiche vertieft werden. Dazu werden alle relevanten Metriken hinsichtlich der Kategorien betrachtet. Die folgenden Grafiken werden in dieser Form interpretiert: Die Balken repräsentieren die Mittelwerte für die jeweilige Kategorie. Die Fehlerbalken sind mit einen Standardfehler mit einem Multiplikator von 1 angetragen. Die konkreten Mittelwerte sind angegeben.

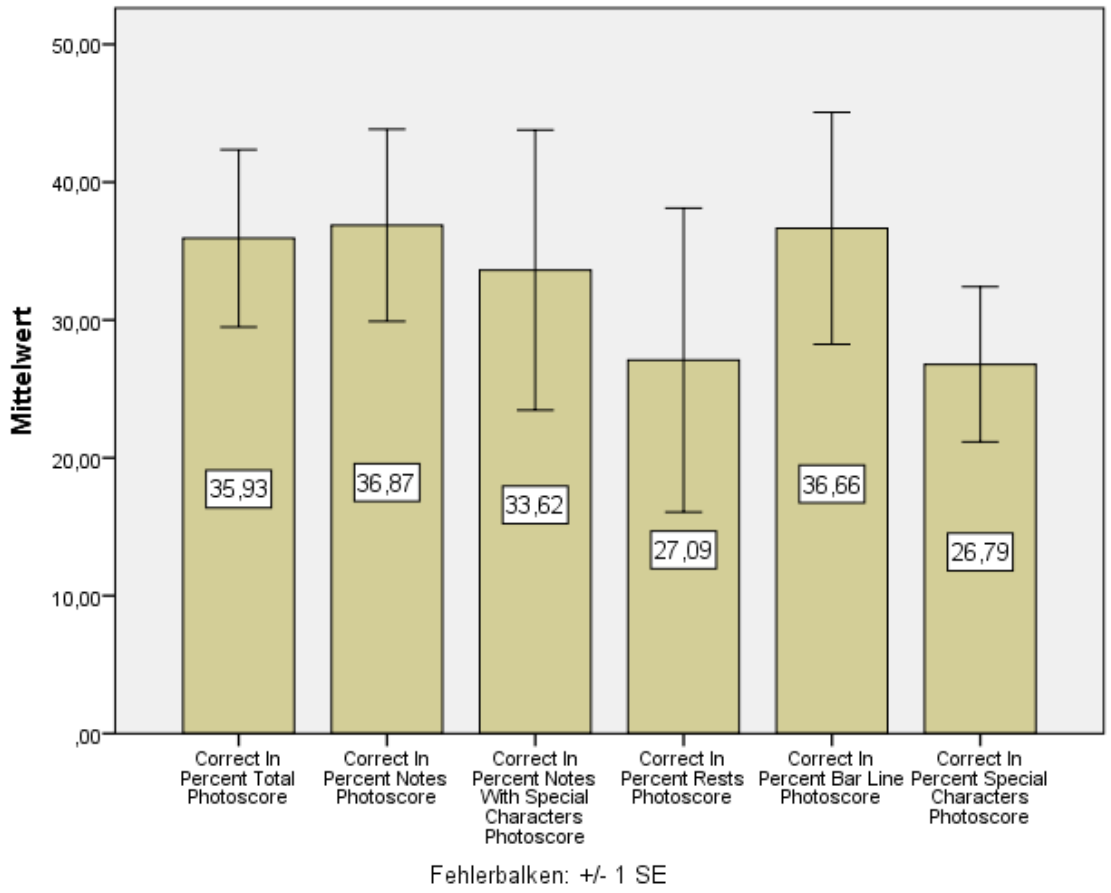


Abbildung 29: Balkendiagramm – Kategorialer Mittelwertvergleich Correct In Percent

Bei der Accuracy zeigt sich, dass Noten, Noten mit Sonderzeichen und Taktstriche gleichmäßig um einen Mittelwert von ca. 35% erkannt werden. Die Erkennungsrate bei Pausen und bei Sonderzeichen (gemäß der Definition dieser Studie) ist um etwa 10% schlechter. Insgesamt sind jedoch alle Erkennungsraten weit unter dem Zielwert von 90%.

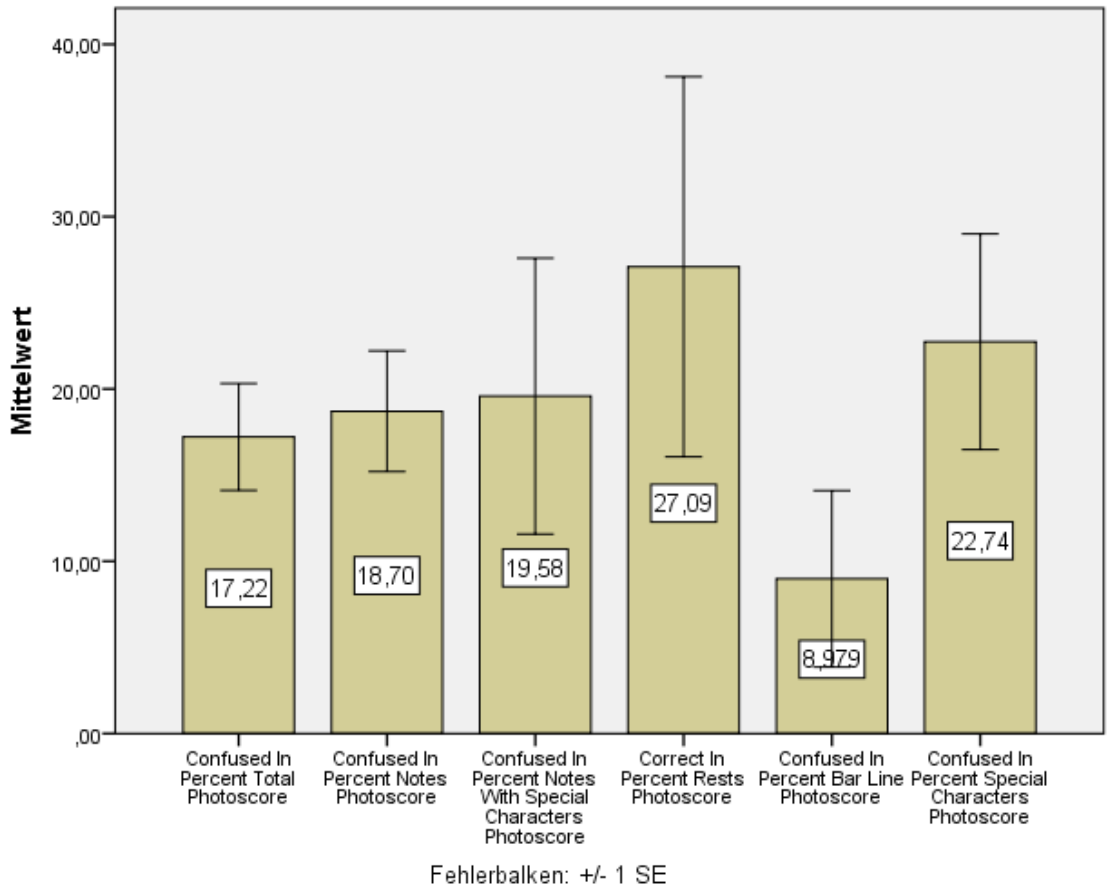


Abbildung 30: Balkendiagramm – Kategorialer Mittelwertvergleich Confused In Percent

Das Ergebnis für die Accuracy führt sich auch bei den Confused-Werten weiter. Besonders Noten haben vergleichsweise gute Raten unter 20%, während Pausen und Sonderzeichen häufiger verwechselt werden. Taktstriche werden am häufigsten korrekt erkannt.

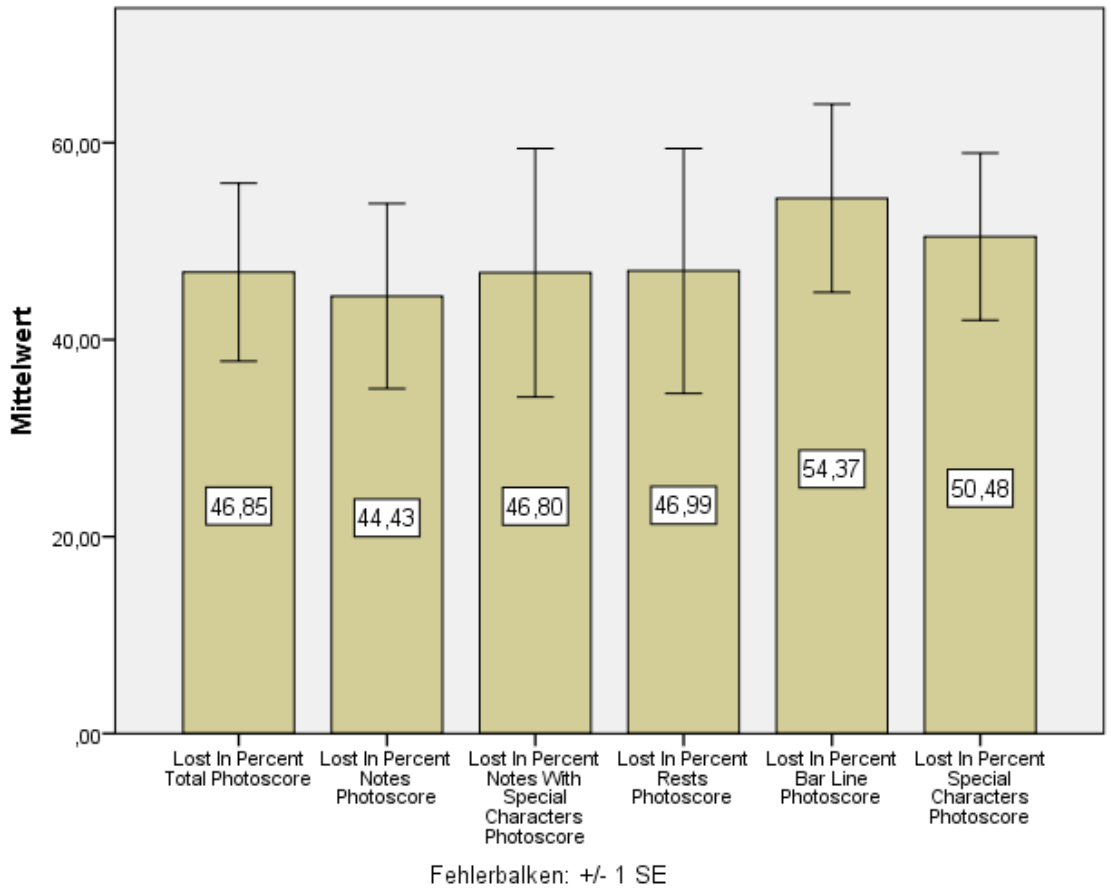


Abbildung 31: Balkendiagramm – Kategorialer Mittelwertvergleich Lost In Percent

Aufgrund der großen Anzahl an von Photoscore gar nicht als Notenblätter erkannten Liedblätter sind die Lost-Werte gleichmäßig hoch, jedoch auch hier bei Noten geringfügig besser als bei Taktstrichen und Sonderzeichen. Dies belegt, dass Taktstriche eher gar nicht erkannt werden als fehlerhaft.

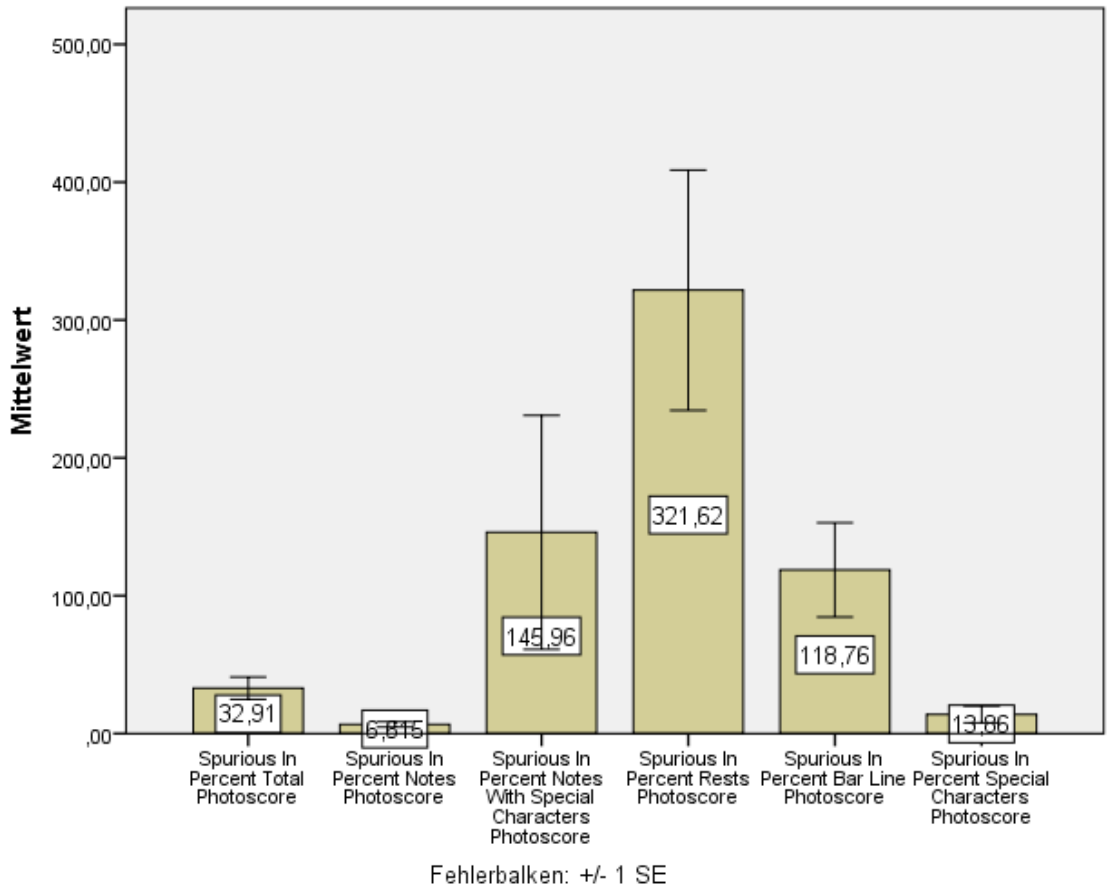


Abbildung 32: Balkendiagramm – Kategorialer Mittelwertvergleich Spurious In Percent

Im Bereich der Spurious-Werte werden schwerwiegende Unterschiede in der Leistung pro Kategorie deutlich. So werden prozentual betrachtet deutlich mehr Pausen und Taktstriche als Noise erzeugt als beispielsweise Noten. Dies liegt unter anderem auch daran, dass Photoscore unvollständige und leere Notenzeilen, wie sie in der Liedblattsammlung punktuell vorkommen mit Pausen und Taktstrichen ersetzt. Gleichzeitig liegen diese Zeichen absolut betrachtet seltener vor, was sie anfällig für starke prozentuale Noise-Werte macht. Mit den absoluten Werten kann man die Grafik als Kreisdiagramm darstellen:

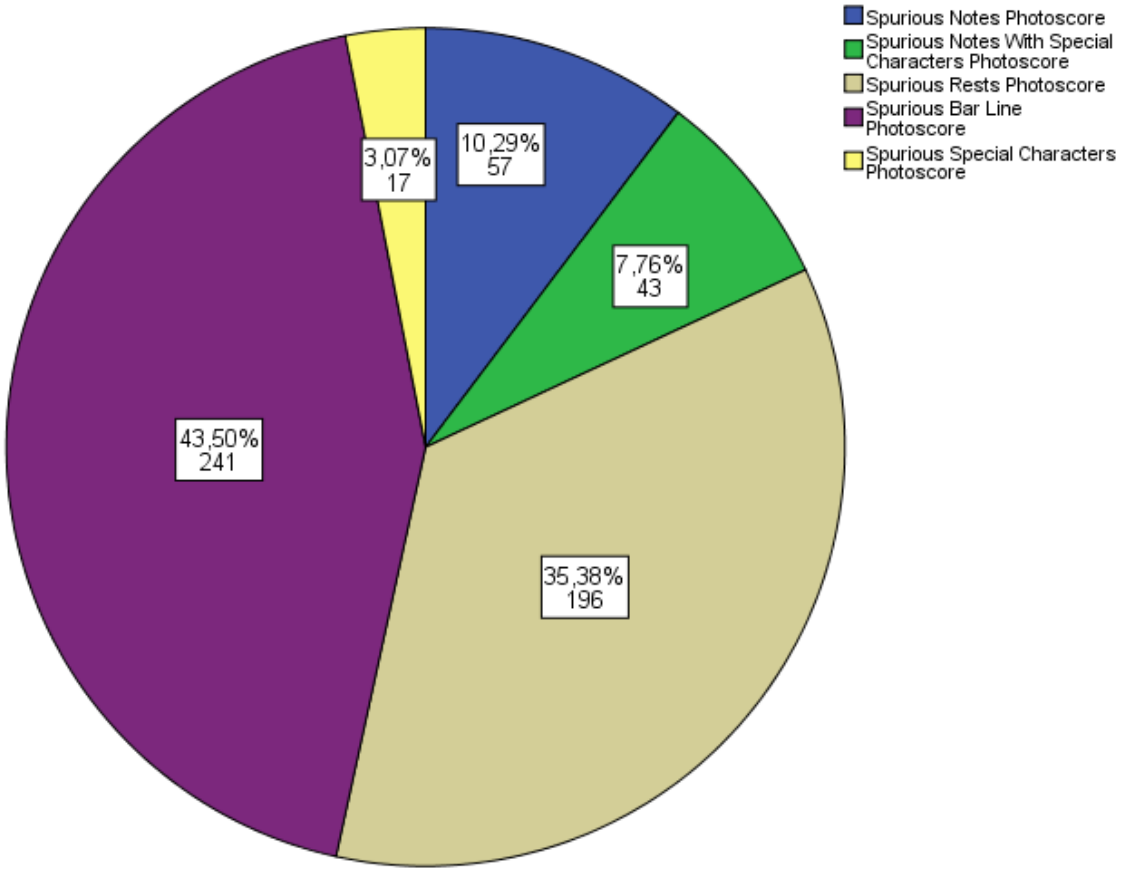


Abbildung 33: Kreisdiagramm – Kategorialer Vergleich Spurious

Dieses Diagramm bestätigt das obige Balkendiagramm auch bei den absoluten Werten. Fast 80% aller Noise sind Taktstriche und Pausen. Noten sind eher selten und Sonderzeichen nur zu 3% als Spurious-Zeichen vorhanden.

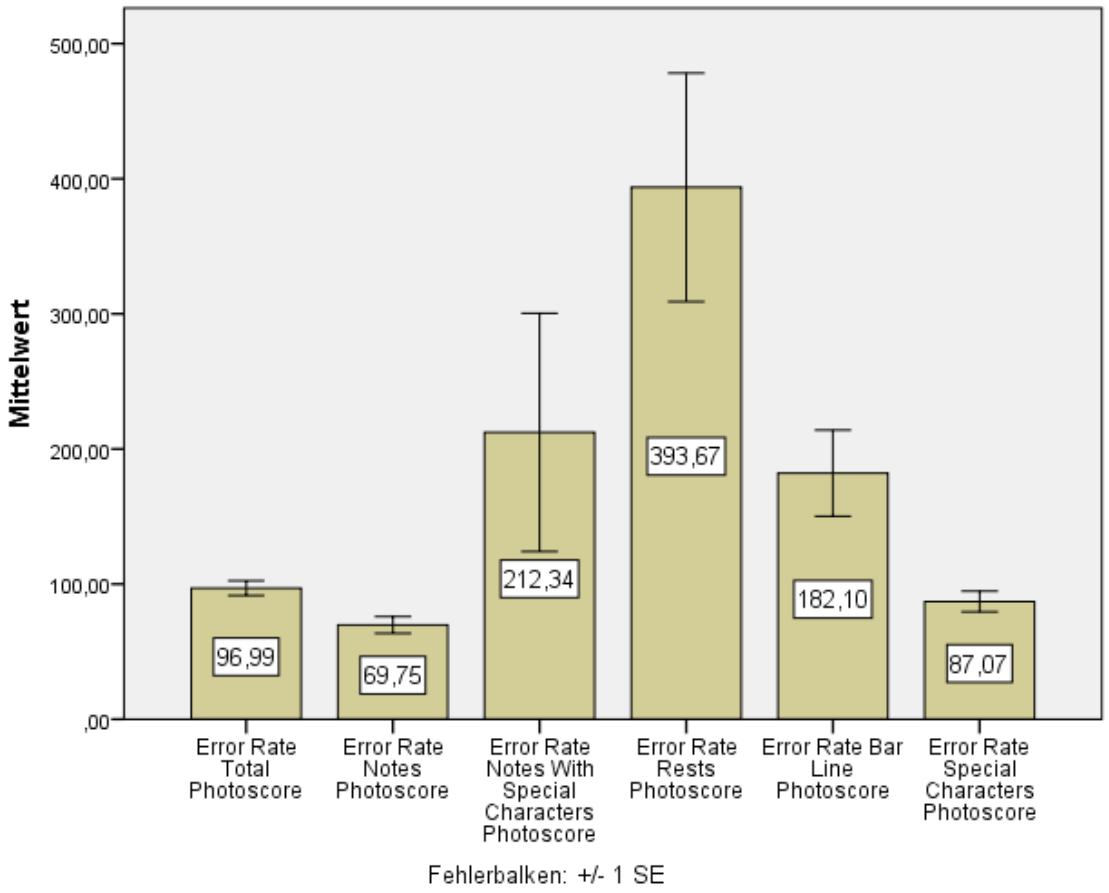


Abbildung 34: Balkendiagramm – Kategorialer Mittelwertvergleich Error Rate

Bei den Fehlerraten als wichtigstes Maß zur Interpretation der OMR-Leistung, erkennt man den Einfluss der hohen Spurious-Anteile für Pausen und Taktstriche. Der höhere Wert bei Noten mit Sonderzeichen liegt vor allem daran, dass diese sehr selten sind. Insgesamt ist die Leistung bei Noten am besten. Fast 70% Fehlerrate bedeuten aber immer noch, dass das OMR einen Output erzeugt von dem mehr als die Hälfte falsch, verloren oder Noise ist.

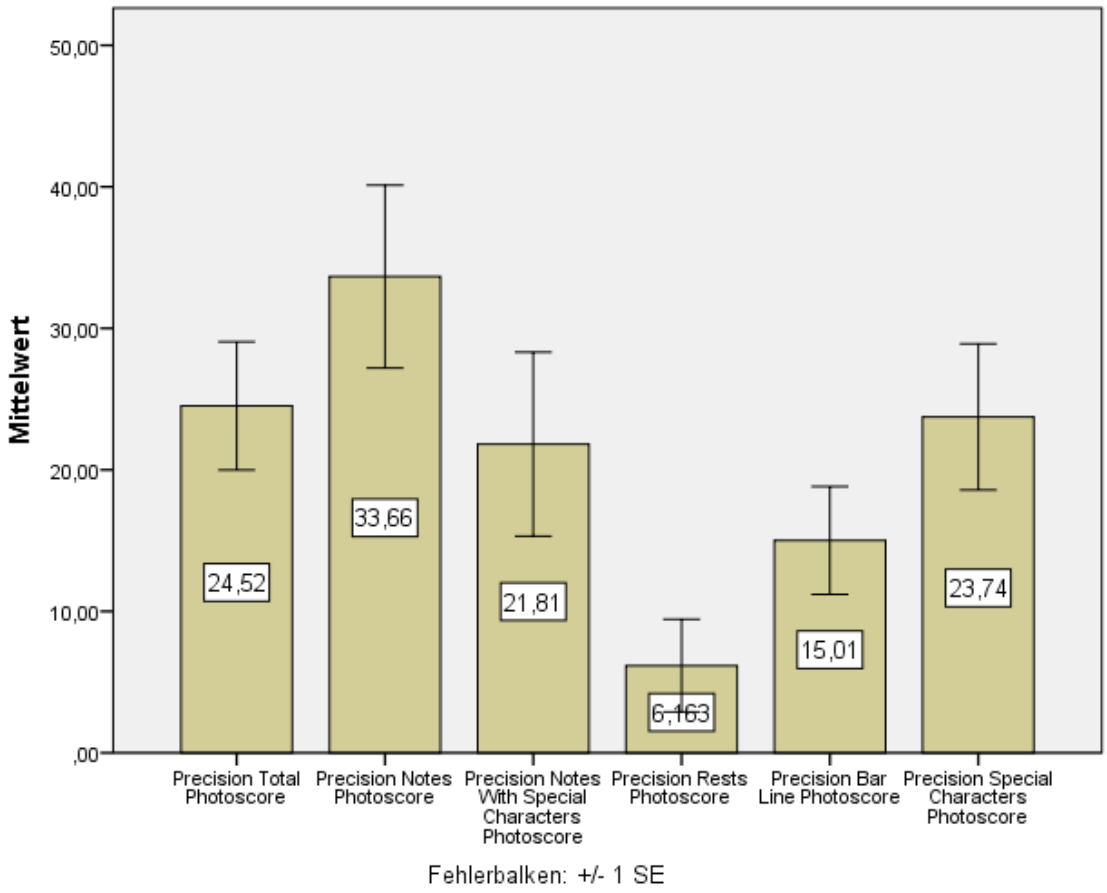


Abbildung 35: Balkendiagramm – Kategorialer Mittelwertvergleich Precision

An der Precision lässt sich nochmals erkennen, dass Noten am besten erkannt werden. Der Mittelwert ist fast gleich der Accuracy, was auf die geringe Noise-Erzeugung weist. Diese Noise-Erzeugung sorgt für sehr schlechte Precision-Werte bei Pausen und Taktstriche. Im Falle von Pausen bedeutet der angegebene Wert, dass nur 5% aller Pausen, die das OMR erzeugt (bezogen auf die Gesamtzahl im Original) korrekt sind. Alle Kategorien liefern Ergebnisse weit unter wünschenswerten Werten (>90%).

3.1.3 SharpEye

Die deskriptive Statistik für die Performanz von SharpEye wird mit folgender Tabelle zusammengefasst. Es bezieht sich auf alle Zeichen also kategorie-unabhängig zusammengezählt.

Tabelle 16: Deskriptive Statistik – Total SharpEye

Deskriptive Statistiken

	N	Minimum	Maximum	Mittelwert	Standardabweichung	Varianz
Total General	20	35,00	182,00	89,7000	46,63757	2175,063
Correct In Percent Total SharpEye	20	,00	22,86	4,0798	5,86361	34,382
Confused In Percent Total SharpEye	20	,00	21,57	5,0967	6,18183	38,215
Lost In Percent Total SharpEye	20	71,57	100,00	90,8234	9,13140	83,382
Spurios In Percent Total SharpEye	20	,00	1,96	,2199	,51035	,260
ErrorRate Total SharpEye	20	77,14	100,00	96,1401	5,74479	33,003
Precision Total SharpEye	20	,00	22,86	4,0633	5,84648	34,181
Gültige Anzahl (listenweise)	20					

Die Erkennungsrate liegt bei einem Mittelwert von 5%. Die beste Erkennungsrate beträgt knapp 23%. Der häufigste Fehler ist dabei Lost. So gehen im Schnitt 90% aller Zeichen verloren. Dementsprechend sind die anderen Fehlerarten zu vernachlässigen und die Error Rate fast gleich der Lost-Rate. Eine große Zahl von Liedblättern wird fast überhaupt nicht erkannt. Dies zeigt auch das Histogramm:

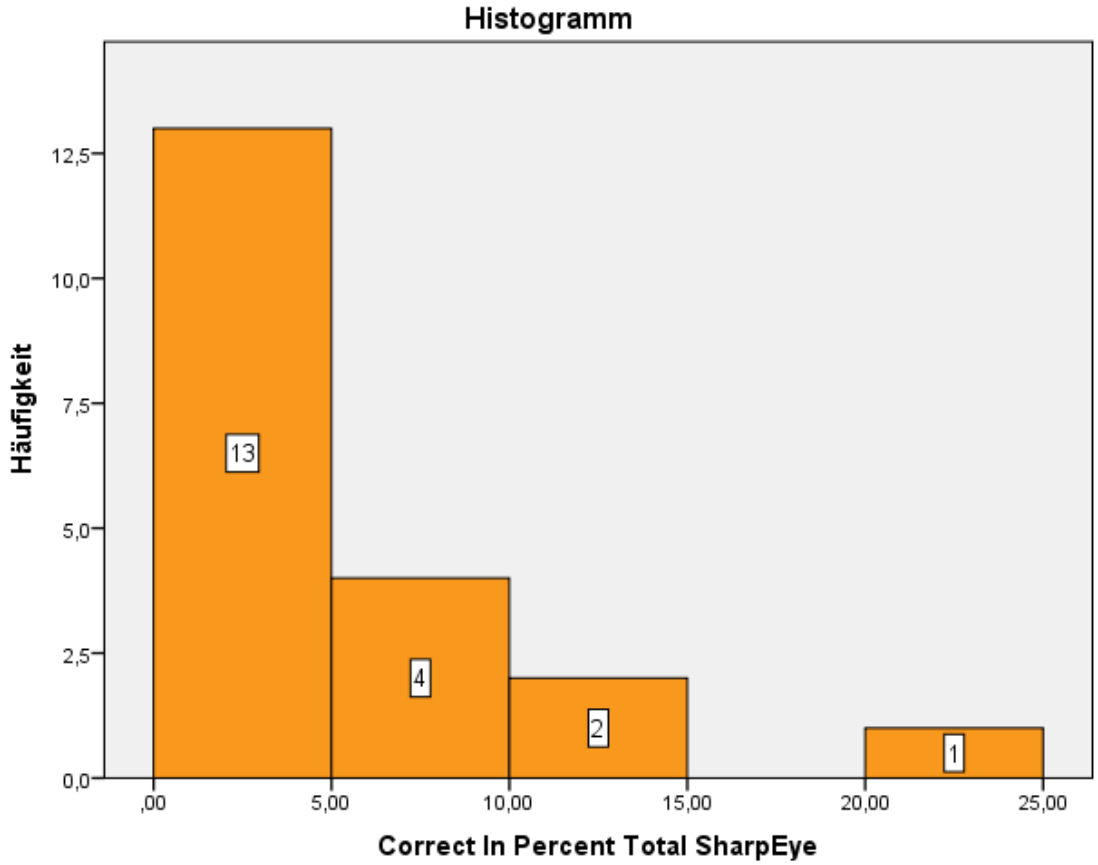


Abbildung 36: Histogramm – Correct In Percent Total SharpEye

Auch in der absoluten Verteilung der Typen, visualisiert in einem Kreisdiagramm werden die großen Lost-Raten deutlich:

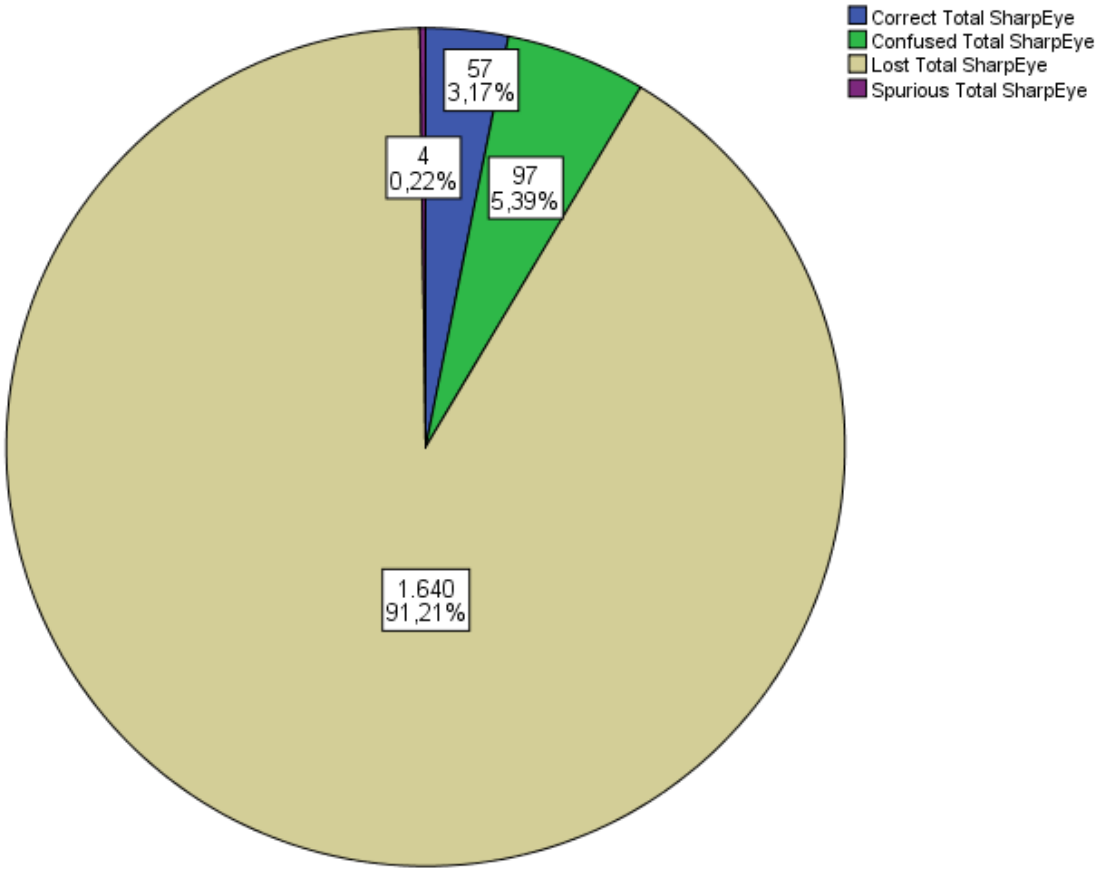


Abbildung 37: Kreisdiagramm – Typvergleich Total SharpEye

Es werden insgesamt nur 57 Zeichen korrekt erkannt. Dies findet kategorie-unabhängig so statt. Zur genaueren Betrachtung seien hier jedoch noch die deskriptiven Werte für die Noten angegeben, welche die bedeutendste Kategorie ausmachen:

Tabelle 17: Deskriptive Statistik – Notes SharpEye

Deskriptive Statistiken

	N	Minimum	Maximum	Mittelwert	Standardabweichung	Varianz
Total Notes	20	24,00	127,00	57,2500	30,84575	951,461
Correct In Percent Notes SharpEye	20	,00	32,00	5,3839	8,84178	78,177
Confused In Percent Notes SharpEye	20	,00	25,71	6,5897	8,07724	65,242
Lost In Percent Notes SharpEye	20	64,00	100,00	88,0264	12,85638	165,287
Spurious In Percent Notes SharpEye	20	,00	2,17	,1745	,55480	,308
ErrorRate Notes SharpEye	20	68,00	100,00	94,7906	8,79178	77,295
Precision Notes SharpEye	20	,00	32,00	5,3692	8,82974	77,964
Gültige Anzahl (listenweise)	20					

Die Ergebnisse weisen auch hier auf eine sehr schlechte Leistung hin. Auch hier wurden die Daten kategorial untersucht. Von einer genaueren Betrachtung wird jedoch abgelassen. Das Tool SharpEye ist basierend auf den Ergebnissen über dem Testkorpus ungeeignet für die Erschließung der Liedblattsammlung. Eine genauere Analyse ist nicht zielführend.

3.1.4 Capella Scan

Nachfolgend die Ergebnisse für die Metriken bei Capella-Scan (kategorie-unabhängig):

Tabelle 18: Deskriptive Statistik – Total Capella Scan

Deskriptive Statistiken

	N	Minimum	Maximum	Mittelwert	Standardab-weichung	Varianz
Total General	20	35,00	182,00	89,7000	46,63757	2175,063
Correct In Percent Total CapellaScan	20	,00	21,57	8,3099	7,44807	55,474
Confused In Percent Total CapellaScan	20	,00	41,18	17,4541	12,22391	149,424
Lost In Percent Total CapellaScan	20	37,25	100,00	74,2359	18,05303	325,912
Spurious In Percent Total CapellaScan	20	,00	17,33	2,6383	4,03549	16,285
Error Rate Total CapellaScan	20	82,35	109,33	94,3284	7,27787	52,967
Precision Total CapellaScan	20	,00	20,75	8,0210	7,22281	52,169
Gültige Anzahl (listenweise)	20					

Auch hier ist das Ergebnis mit einer Erkennungsrate von 8% im Mittel sehr schlecht. Die im Vergleich zu SharpEye geringere Lost-Rate von 75% weist auf eine geringfügig bessere Leistung hin. Die Fehlerrate ist aber nahezu gleich groß und wird hier durch höhere Confused-Werte ergänzt. Sehr hoch ist auch hier die Zahl der nicht erkannten Blätter, wie man am Histogramm für die Accuracy und an der absoluten Typenverteilung im Kreisdiagramm sehen kann:

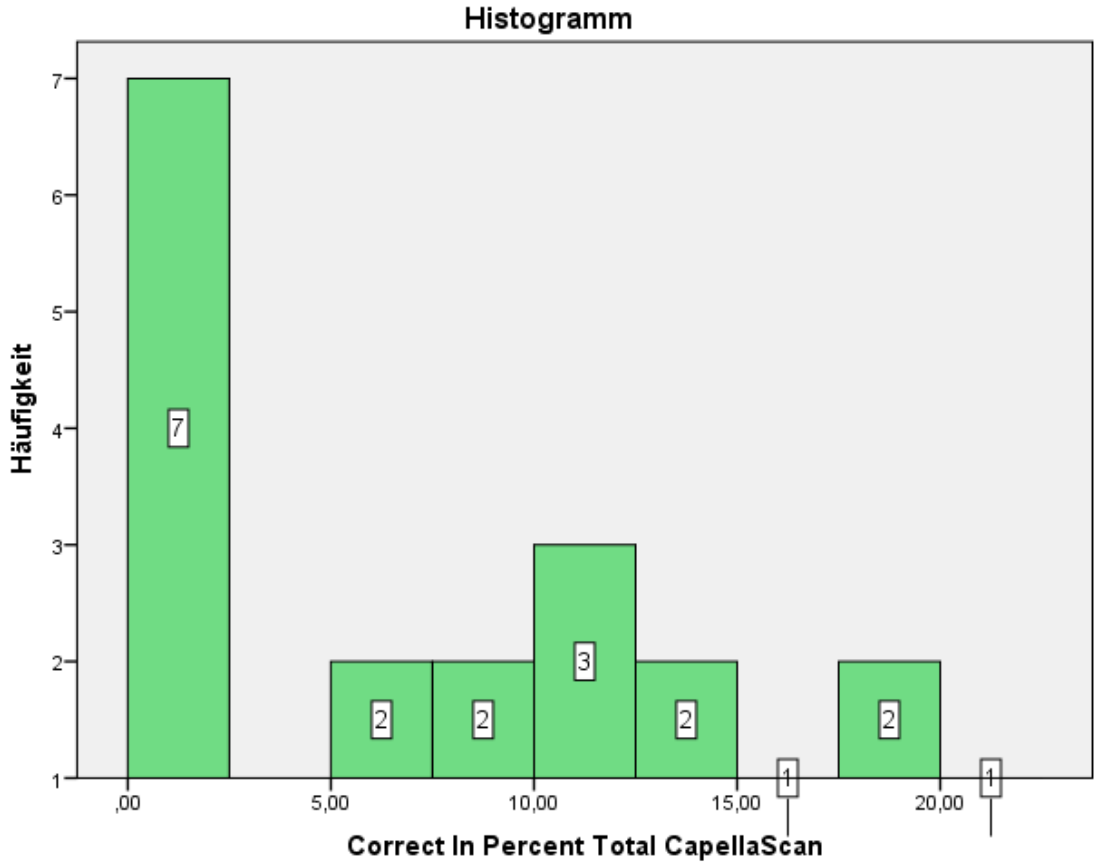


Abbildung 38: Histogramm – Correct In Percent Total Capella Scan

Auch das Maximum von 22% ist weit von einer „guten“ Erkennungsrate von 90% entfernt.

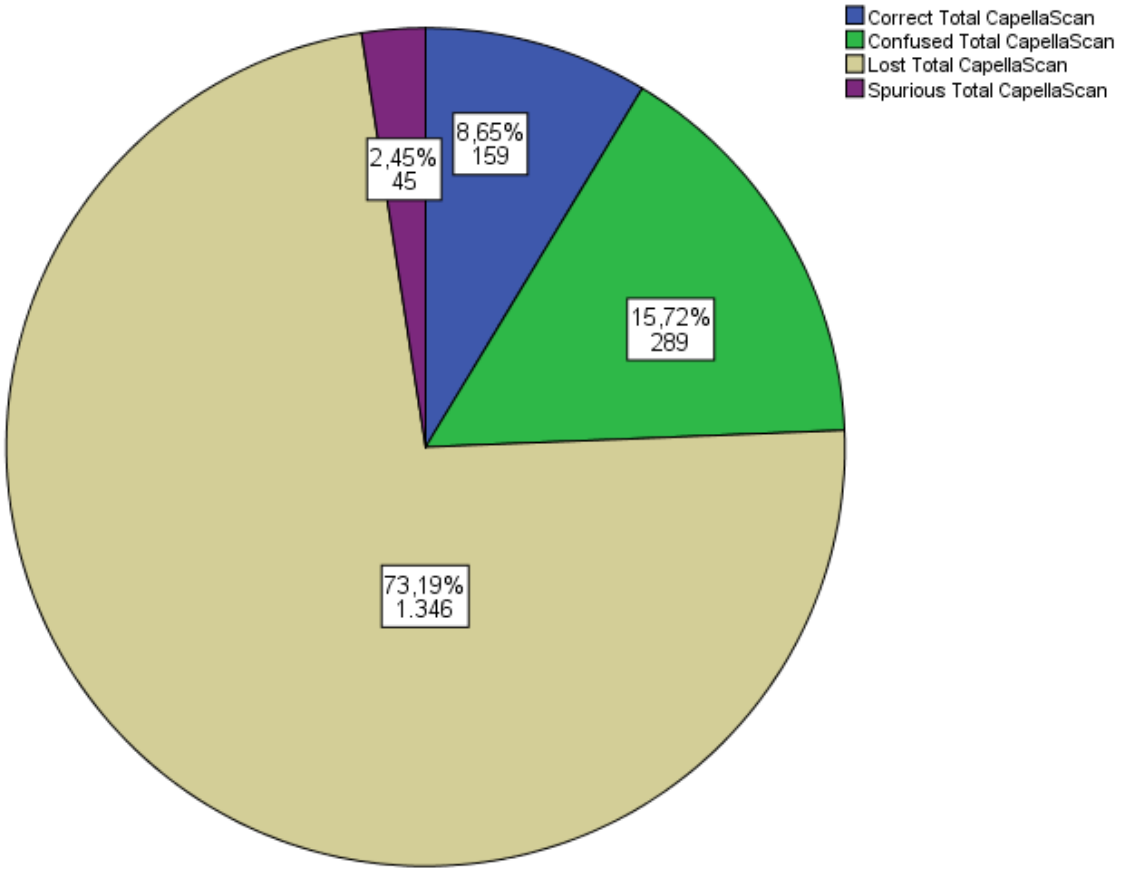


Abbildung 39: Kreisdiagramm – Typvergleich Total Capella Scan

Im Kreisdiagramm wird die große Lost-Rate deutlich, aber auch bei erkannten Zeichen ist die Fehlerrate über eine hohe Confused-Rate sehr groß.

Auch hier fand eine tiefergehende, kategoriale Analyse statt. Diese hier zu präsentieren ist jedoch nicht zielführend. Auch Capella-Scan ist aufgrund der unterdurchschnittlichen Leistung nicht zur Erschließung geeignet. Zur Vervollständigung seien hier noch die deskriptiven Daten allein für die Kategorie Noten angegeben:

Tabelle 19: Deskriptive Statistik – Notes Capella Scan

Deskriptive Statistiken

	N	Minimum	Maximum	Mittelwert	Standardab-weichung	Varianz
Total Notes	20	24,00	127,00	57,2500	30,84575	951,461
Correct In Percent Notes CapellaScan	20	,00	13,04	3,6194	4,69394	22,033
Confused In Percent Notes CapellaScan	20	,00	46,43	17,3156	14,76130	217,896
Lost In Percent Notes Ca-pellaScan	20	46,43	100,00	79,0650	16,47788	271,521
Spurious In Percent Notes CapellaScan	20	,00	7,14	1,7892	2,67695	7,166
Error Rate Notes Capel-laScan	20	89,81	105,49	98,1699	4,62982	21,435
Precision Notes Capel-laScan	20	,00	12,24	3,5230	4,55340	20,733
Gültige Anzahl (listen-weise)	20					

3.2 Inferenzstatistik und Vergleich

3.2.1 Hypothesenbildung

Eine Frage, die mit der vorliegenden Studie beantwortet werden soll, ist, welches Tool am besten für die Erschließung des Datenbestands geeignet ist. Die deskriptive Statistik liefert eine eindeutige Antwort. Dennoch soll hier mit der hypothesengeleiteten Inferenzstatistik die Antwort auf diese Frage genauer untersucht und beantwortet werden. Dazu werden notwendige Hypothesen gebildet. Die Haupthypothese ist wie folgt formuliert:

H1: *Es gibt Unterschiede in der Leistung der drei Tools (Photoscore, SharpEye, Capella Scan) für den Testkorpus.*

Die Leistung wird dabei variabel über den jeweils betrachteten Performanz-Parameter operationalisiert. Die Hypothese kann man nach Einsicht der deskriptiven Statistik noch in kleinteiligere Einzelhypothesen zerlegen, die eine genaue Richtung enthalten:

H1.1: *Photoscore hat eine bessere Performanz als SharpEye für den Testkorpus.*

H1.2: *Photoscore hat eine bessere Performanz als Capella-Scan für den Testkorpus.*

Zur Beantwortung dieser Fragestellungen über statistische Tests ist eine Nullhypothese notwendig (Leonhart, 2013, S. 177):

H0: *Es gibt keinen Unterschied in der Performanz der OMR-Tools Photoscore, SharpEye, und Capella Scan für den Testkorpus.*

Die Nullhypothese soll dabei zuerst widerlegt werden.

Ähnlich zur Studie im OCR zerfallen die obigen Hypothesen für jede Performanz-Variable noch mal in weitere Einzelhypothese, mit der entsprechenden Richtung für die Variable. Die konkreten Hypothesen verlaufen demnach nach folgendem Verfahren:

H1.1.X (Correct In Percent Total): *Photoscore hat eine bessere Erkennungsrate als SharpEye für den Test-Korpus.*

H1.1.X (Error Rate): *Photoscore hat eine geringere Error Rate als Capella Scan für den Testkorpus.*

Usw.

Aufgrund der eindeutigen Datenlage beschränkt man sich bei der Auswertung auf Correct in Percent, die Error Rate und Lost In Percent, jeweils gesamt. Accuracy und Error Rate sind die zentralen Metriken des OCR. Die Lost-Rate hat nach Einsicht der deskriptiven Statistik eine besondere Bedeutung. Die weiteren Variablen wurden auch analysiert, liefern jedoch keinen Mehrwert. Im Folgenden wird sich immer auf die H1, H1.1 und H1.2 im Kontext der betrachteten Variable bezogen.

3.2.2 Signifikanztests - Statistisches Vorgehen

Bei den Stichproben für die zu untersuchenden Variablen handelt es sich um gepaarte Stichproben (Leonhart, 2013, S. 190). Eine paarweise Zuordnung ist möglich, da jede Variable Werte für jedes einzelne Liedblatt erzeugt. Die Werte der Variablen unterscheiden sich dadurch, dass sie für unterschiedliche Tools erhoben wurden. Tool ist demnach der Innersubjekt-Faktor. Die Signifikanz sowie die Stärke dieses Faktors sollen im Folgenden untersucht werden. Eine Studie im OCR geht nach einem ähnlichen Verfahren vor (Kanungo et al., 1998). Leonhart (2013, S. 486) empfiehlt für diesen Fall die Varianzanalyse mit Messwiederholung für mindestens intervallskalierte Variablen, jedoch werden die Voraussetzungen für diese nicht erfüllt. Zwar liegt die Skalierung mit metrischen Variablen vor, allerdings liegt für die zu analysierenden Variablen keine Normalverteilung vor. Außerdem ist die Stichprobe mit 20 Liedblättern weit unter den Mindest-Emp-

fehlungen (N > 30; Lüpsen, 2015, S. 21). Deswegen wird auf ein schwächeres Testverfahren für ordinalskalierte Variablen ausgewichen: Den Friedman-Test (Leonhart, 2010, S. 177). Dieser transformiert die Daten in eine Rang-Skalierung. Er ist konservativer, verlangt aber weniger Voraussetzungen. Auch ist er besser für Stichproben von geringer Anzahl, wie bei der vorliegenden Studie, geeignet. Mit ihm wird die H1 überprüft. Das Signifikanzniveau beträgt 0.05. Die H1.1 und H1.2, also die paarweisen Unterschiede werden gemäß Leonhart (2010, S. 172) mit dem Wilcoxon-Test untersucht. Dieser muss demnach mehrfach für mehrere Durchgänge durchgeführt werden. Dadurch steigt die Wahrscheinlichkeit eines α-Fehlers und eine Bonferroni-Korrektur wird notwendig (Bortz, 2005, S. 129). Dazu wird das Signifikanzniveau durch die Zahl der Tests geteilt, also hier 0.05/2 = 0.025. Dieses Niveau gilt also für die paarweisen Tests. Es gibt unterschiedliche Korrekturverfahren, jedoch ist dieses das konservativste, die Wahrscheinlichkeit die H1.1 oder H1.2 also fälschlicherweise anzunehmen ist eher gering.

Die Ergebnisse werden jeweils mit einer Boxplot-Grafik verdeutlicht. Mehr Informationen zu verschiedenen statistischen Vorgängen findet man im entsprechenden Abschnitt (Kapitel 3) des OCR-Berichts.

3.2.3 Correct In Percent (Accuracy)

Es werden nun die Hypothesen bezüglich der Erkennungsrate untersucht. Zunächst sei hierfür die deskriptive Statistik kurz wiederholt:

Tabelle 20: Deskriptive Statistiken – Correct In Percent

Deskriptive Statistiken

	H	Mittelwert	Standardabweichung	Minimum	Maximum
Correct In Percent Total Photoscore	20	35,9252	28,78305	,00	75,53
Correct In Percent Total SharpEye	20	4,0798	5,86361	,00	22,86
Correct In Percent Total CapellaScan	20	8,3099	7,44807	,00	21,57

Die deskriptive Statistik verdeutlicht nochmal die kaum vorhanden Erkennungsrate von SharpEye und Capella-Scan. Der Signifikanztest (Friedman-Test) bestätigt zunächst die H1:

Tabelle 21: Friedman-Test – Correct In Percent

Teststatistiken[a]

H	20
Chi-Quadrat	14,381
df	2
Asymp. Sig.	,001

a. Friedman-Test

Der Unterschied zwischen den drei Tools ist also hoch signifikant. Der Wilcoxon-Test soll nun überprüfen ob auch die gerichteten Hypothesen H1.1 und H1.2 angenommen werden können:

Tabelle 22: Rangvergleich – Correct In Percent

Ränge

		H	Mittlerer Rang	Summe der Ränge
Correct In Percent Total SharpEye - Correct In Percent Total Photoscore	Negative Ränge	13[a]	9,00	117,00
	Positive Ränge	2[b]	1,50	3,00
	Bindungen	5[c]		
	Gesamtsumme	20		
Correct In Percent Total CapellaScan - Correct In Percent Total Photoscore	Negative Ränge	14[d]	9,36	131,00
	Positive Ränge	2[e]	2,50	5,00
	Bindungen	4[f]		
	Gesamtsumme	20		

a. Correct In Percent Total SharpEye < Correct In Percent Total Photoscore
b. Correct In Percent Total SharpEye > Correct In Percent Total Photoscore
c. Correct In Percent Total SharpEye = Correct In Percent Total Photoscore
d. Correct In Percent Total CapellaScan < Correct In Percent Total Photoscore
e. Correct In Percent Total CapellaScan > Correct In Percent Total Photoscore
f. Correct In Percent Total CapellaScan = Correct In Percent Total Photoscore

Zunächst kann man das Rangverhalten betrachten. Dabei sieht man, dass die Fälle a und d am häufigsten auftreten. Dies sind die Fälle, die die Hypothesen bestätigen. Die Bindungen betreffen vor allem Blätter die gar nicht erkannt wurden. Anschließend kann man einsehen ob die Fälle a und d nun auch jeweils signifikant sind.

Tabelle 23: Wilcoxon-Tests – Correct In Percent

Teststatistiken[a]

	Correct In Percent Total SharpEye - Correct In Percent Total Photoscore	Correct In Percent Total CapellaScan - Correct In Percent Total Photoscore
U	-3,237[b]	-3,258[b]
Asymp. Sig. (2-seitig)	,001	,001

a. Wilcoxon-Test

b. Basierend auf positiven Rängen.

Auch hier wird das strengere Signifikanzniveau von 0.025 unterboten. Die H1.1 und H1.2 können angenommen werden. Photoscore ist bezüglich der Erkennungsrate das bessere Tool. Dennoch sei hier anzumerken das auch Photoscore weit unter sonstigen Standards im OCR/OMR operiert (Holley, 2009).

Die folgende Boxplot-Grafik visualisiert das Ergebnis. Mehr Informationen zu Box-Plots findet man in Kapitel 3.5.4.2 im OCR-Bericht:

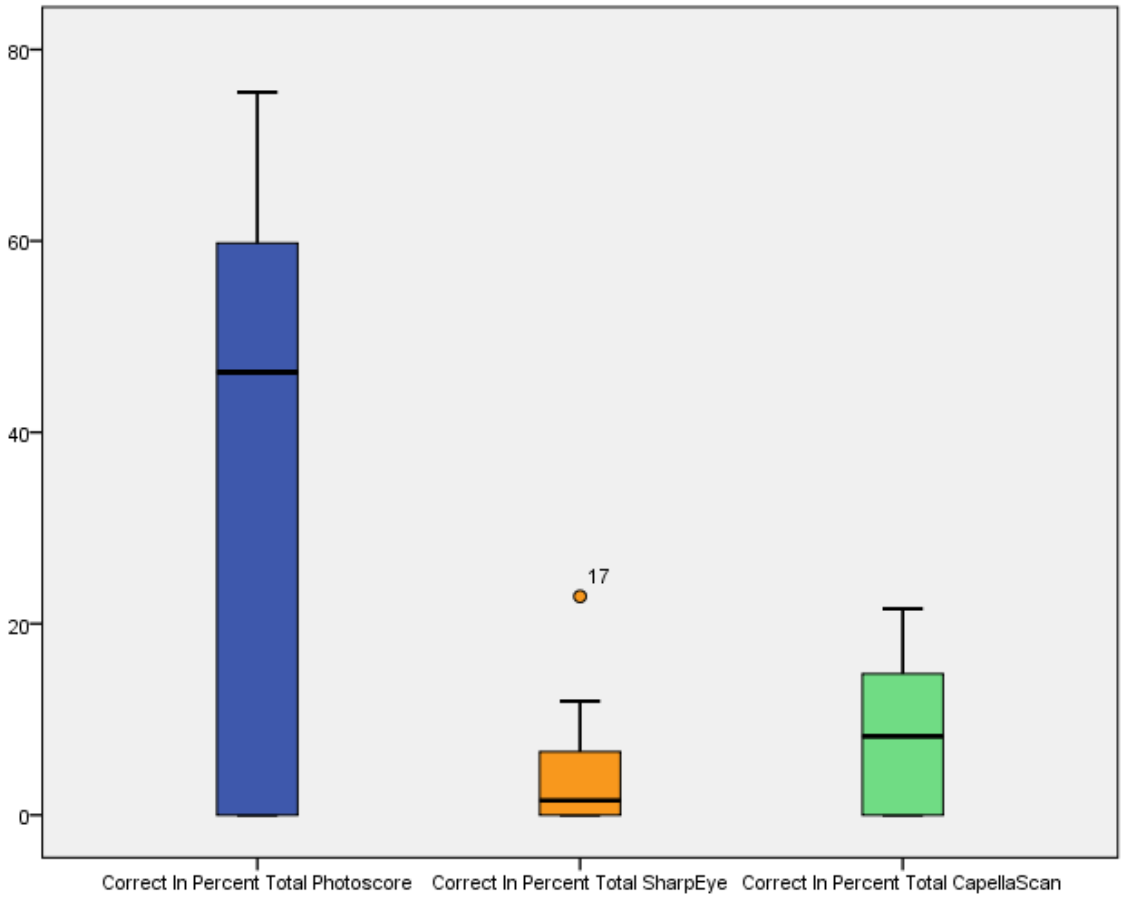


Abbildung 40: Boxplot-Grafik – Correct In Percent

Man sieht deutlich, dass nur Photoscore überhaupt rentable Erkennungsraten über 50% erreicht. Dennoch ist die Streuung stark und bei der Hälfte aller Blätter liegt die Erkennungsrate bei unter 40%. Die Erkennungsraten für die anderen beiden Tools liegen relativ gleichmäßig verteilt für alle Liedblätter unter 20%. Ferner sei anzumerken, dass das Ergebnis für die Hauptkategorie Noten analog ist, was daran liegt, dass die Mehrzahl der Zeichen Noten sind.

3.2.4 Lost In Percent

Die deskriptive Statistik für Lost In Percent ist eindeutig:

Tabelle 24: Deskriptive Statistiken – Lost In Percent

Deskriptive Statistiken

	H	Mittelwert	Standardabweichung	Minimum	Maximum
Lost In Percent Total Photoscore	20	46,8545	40,45758	4,26	100,00
Lost In Percent Total SharpEye	20	90,8234	9,13140	71,57	100,00
Lost In Percent Total CapellaScan	20	74,2359	18,05303	37,25	100,00

Sowohl SharpEye als auch Capella Scan „verlieren“ fast den gesamten Datenbestand. Auch bei Photoscore ist die Rate hoch, die größere Standardabweichung weist jedoch auf insgesamt bessere Leistung hin. Zunächst folgt das Ergebnis des Friedman-Tests:

Tabelle 25: Friedman-Test – Lost In Percent

Teststatistiken[a]

H	20
Chi-Quadrat	20,375
df	2
Asymp. Sig.	,000

a. Friedman-Test

Die H1 kann angenommen werden. Die Wilcoxon-Tests liefern folgendendes Ergebnis:

Tabelle 26: Rangvergleiche – Lost In Percent

Ränge

		H	Mittlerer Rang	Summe der Ränge
Lost In Percent Total SharpEye - Lost In Percent Total Photoscore	Negative Ränge	2[a]	2,50	5,00
	Positive Ränge	14[b]	9,36	131,00
	Bindungen	4[c]		
	Gesamtsumme	20		
Lost In Percent Total CapellaScan - Lost In Percent Total Photoscore	Negative Ränge	3[d]	3,67	11,00
	Positive Ränge	13[e]	9,62	125,00
	Bindungen	4[f]		
	Gesamtsumme	20		

a. Lost In Percent Total SharpEye < Lost In Percent Total Photoscore

b. Lost In Percent Total SharpEye > Lost In Percent Total Photoscore

c. Lost In Percent Total SharpEye = Lost In Percent Total Photoscore

d. Lost In Percent Total CapellaScan < Lost In Percent Total Photoscore

e. Lost In Percent Total CapellaScan > Lost In Percent Total Photoscore

f. Lost In Percent Total CapellaScan = Lost In Percent Total Photoscore

Die Hypothesen werden durch die Fälle b und e repräsentiert. Diese sind am häufigsten.

Tabelle 27: Wilcoxon-Tests – Lost In Percent

Teststatistiken[a]

	Lost In Percent Total SharpEye - Lost In Percent Total Photoscore	Lost In Percent Total CapellaScan - Lost In Percent Total Photoscore
U	-3,258[b]	-2,947[b]
Asymp. Sig. (2-seitig)	,001	,003

a. Wilcoxon-Test

b. Basierend auf negativen Rängen.

Unter dem Signifikanzniveau 0.025 kann die H1.1 angenommen werden für Lost In Percent. Photoscore hat also eine bessere Lost-Rate als SharpEye. Diese Aussage ist trivial, da SharpEye fast alle Zeichen nicht erkennt. Für den Vergleich mit Capella Scan ist der Vergleich jedoch nicht signifikant. Die H1.2 muss hier abgelehnt werden. Dies relativiert die Leistung von Photoscore. Diese Relativierung wird vor allem beim nächsten Parameter Error Rate deutlich.

Die Boxplot-Grafik verdeutlicht noch das Ergebnis:

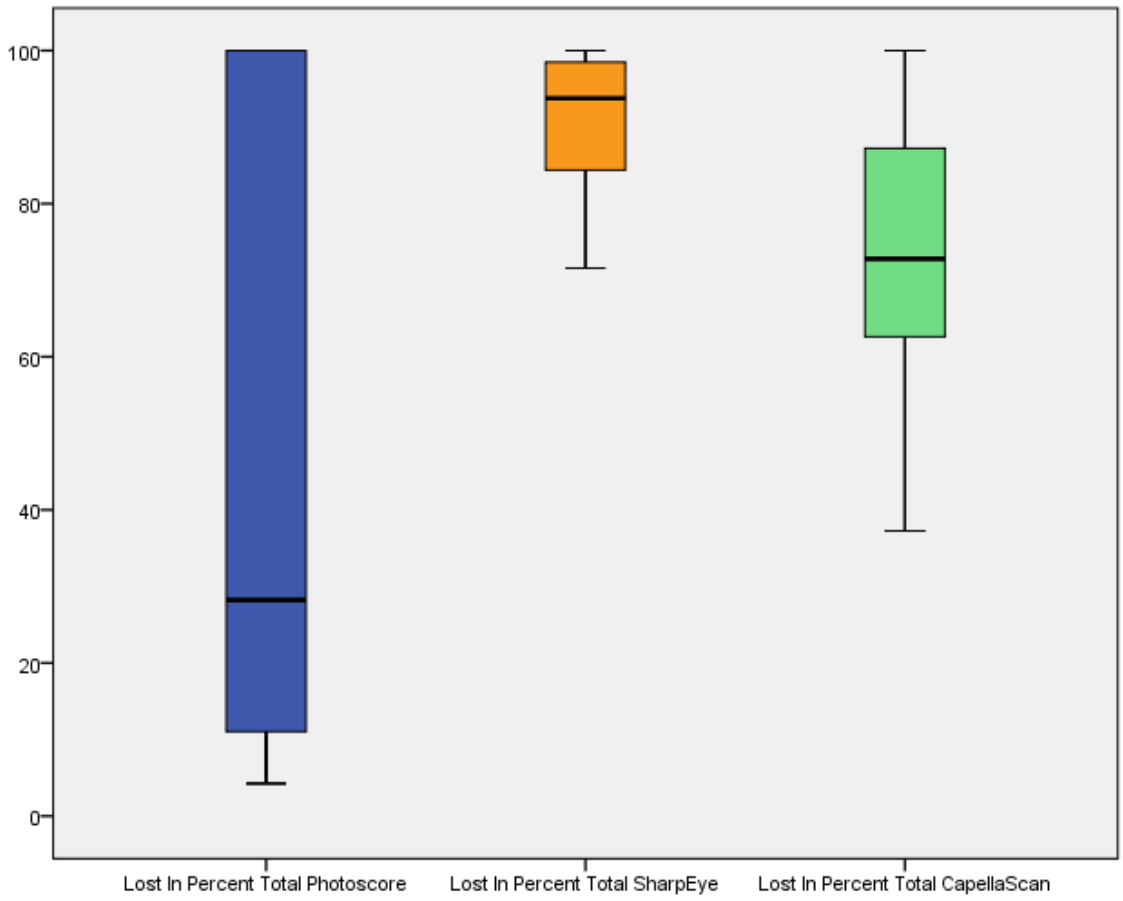


Abbildung 41: Boxplot-Grafik – Lost In Percent

Auffällig ist bei SharpEye, dass alle Blätter eine Lost-Rate über 70% haben. Die Lost-Rate von Capell- Scan streut relativ gleichmäßig und normalverteilt zwischen 40% und 100%. Bei Photoscore ist die starke Varianz zwischen 0 und 100% deutlich. Dennoch haben die Hälfte aller Blätter eine Lost-Rate von unter 30%. Dies ist in etwa der Median. Zwischen den Blättern die zu 30% und zu 100% eine Lost-Rate haben ist die Streuung sehr gleichmäßig. Insgesamt weist die Varianz auf die zwar bessere Leistung als die anderen Tools hin, bestätigt jedoch auch, dass diese Leistung eher durchwachsen ist.

3.2.5 Error Rate

Bei der Error Rate weist die Statistik auf eine Ablehnung der Hypothese:

Tabelle 28: Deskriptive Statistiken – Error Rate

Deskriptive Statistiken

	H	Mittelwert	Standardabweichung	Minimum	Maximum
Error Rate Total Photoscore	20	96,9869	24,63313	37,23	150,00
ErrorRate Total SharpEye	20	96,1401	5,74479	77,14	100,00
Error Rate Total CapellaScan	20	94,3284	7,27787	82,35	109,33

Tatsächlich ist die Error Rate annähernd gleich. Bei Photoscore liegt das an großen Spurios-Werten für Taktstriche und Pausen. Bei den anderen Tools ist vor allem die Lost-Rate sehr hoch, wie schon gezeigt wurde. Der Friedman-Test weist dementsprechend keine Signifikanz nach:

Tabelle 29: Friedman-Test– Error Rate

Teststatistiken[a]

H	20
Chi-Quadrat	2,066
df	2
Asymp. Sig.	,356

a. Friedman-Test

Die H1 wird abgelehnt und die H0 für die Variable Error Rate angenommen. Von paarweisen Vergleichen kann abgesehen werden. Die Error Rate ist der zentrale Parameter im OMR/OCR. Dies soll jedoch nicht bedeuten, dass zwischen den Tools keine Unterschiede bestehen. Die relevantesten wurden schon nachgewiesen. Vielmehr soll gezeigt werden, dass obwohl das Tool Photoscore eine bessere Erkennungsrate hat, die Error Rate darauf hinweist, dass die Fehlererzeugung äquivalent dazu ist, nichts zu erkennen, wie eben die anderen beiden Tools. Auch Photoscore ist also weit unter notwendigen Leistungs-Maßstäben. Die Boxplot-Grafik verdeutlicht die gleichmäßig schwache Leistung aller Tools in Bezug auf die allgemeine Fehlerrate:

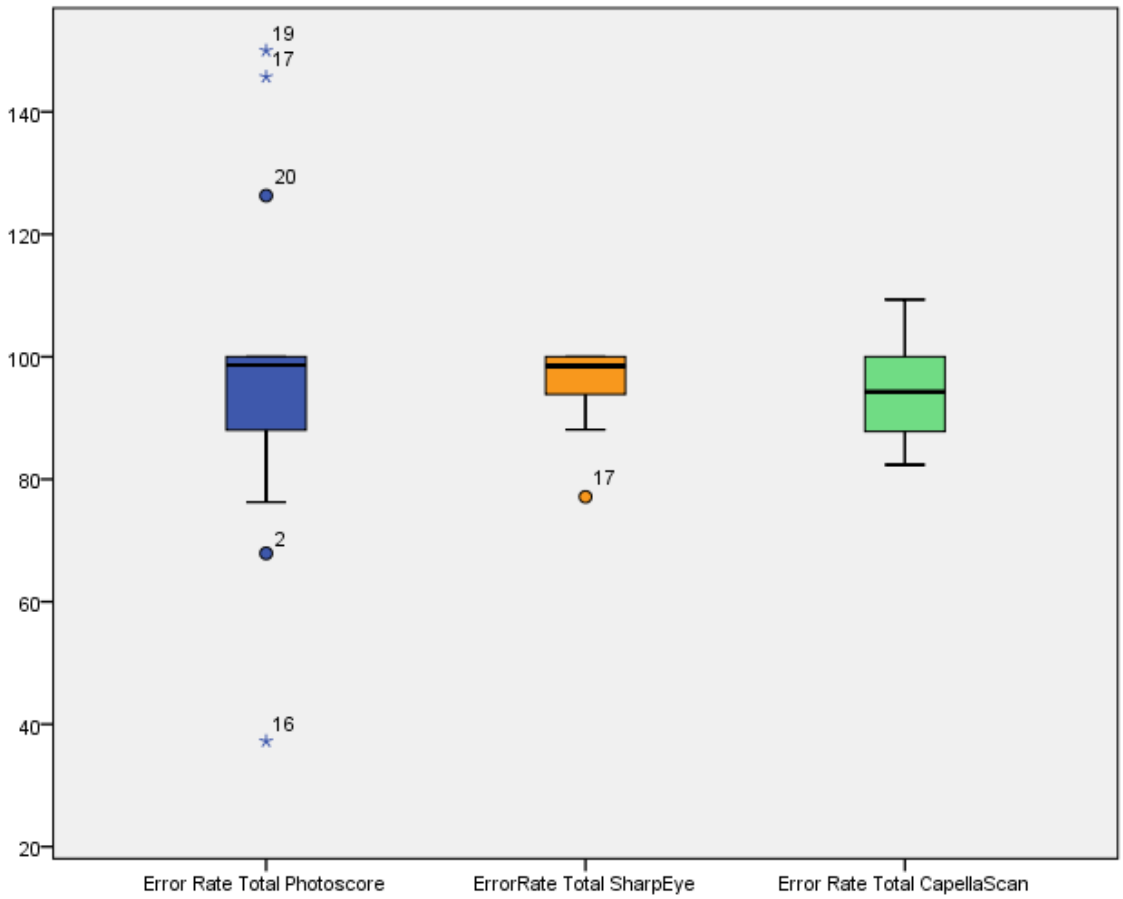


Abbildung 42: Boxplot-Grafik – Error Rate

Der weiter nach unten reichende untere Whisker bei Photoscore, als auch vereinzelte Ausreißer und Extemwerte weisen auf eine geringfügig bessere Leistung als die anderen Tools hin. Bei Capella-Scan ist die Streuung etwas stärker. Insgesamt sind die Boxplots jedoch weitestgehend gleich im Bereich zwichen 90 und 100%. Dies bedeutet dass im Bezug zur Zeichenanzahl des Originals, genauso viele Fehler produziert werden und Zeichen ausgebessert, hinzugefügt oder gelöscht werden müssen.

4 Diskussion

In der Diskussion sollen die Ergebnisse der Evaluation behandelt werden und welche Schlussfolgerungen daraus resultieren, ob also eine Tool-Nutzung sinnvoll ist für die Erschließung der Liedblattsammlung.

Insgesamt sind die Erkennungsraten aller drei evaluierten Tools zu gering, um einen Digitalisierungsprozess der Liedblattsammlung mithilfe von OMR effizient zu gestalten. Über Inferenzstatistik konnte gezeigt werden, dass Photoscore für die Erkennungsraten

ein signifikant besseres Tool ist als SharpEye und Capella-Scan. Dies kann an der Handschriften-Funktion von Photoscore liegen. Bezüglich der Fehlerrate konnte diese Überlegenheit jedoch nicht nachgewiesen werden, was beweist, dass auch Photoscore sehr große Schwächen bei der Erkennung hat. Gemäß Empfehlungen von Holley (2009) zur minimal, notwendigen Erkennungsrate muss man alle drei Tools verwerfen. In der Studie von Bellini, Bruno und Nesi (2007) haben die verwendeten Tools deutlich bessere Erkennungsraten (ca. 96% bei Noten). Es ist naheliegend, dass der große Unterschied durch die Beschaffenheit des Korpus entsteht. Bei der genannten Studie wurden normale gedruckte Notenblätter verwendet. Holley (2009) konnte im OCR feststellen, dass die Leistung bei historischen Dokumenten deutlich unter 90% liegen kann. Dies sagt gemäß Holley weniger über die Tools als über die Datengrundlage aus.

Als Hauptgrund für die schlechten Ergebnisse ist die handschriftliche Notierung der Noten anzusehen. Einen Beleg dazu gibt der Vergleich zwischen drei Notenblättern mit unterschiedlicher Handschrift in Zusammenhang mit den großen, daraus resultierenden Differenzen in der Erkennungsrate von Photoscore.

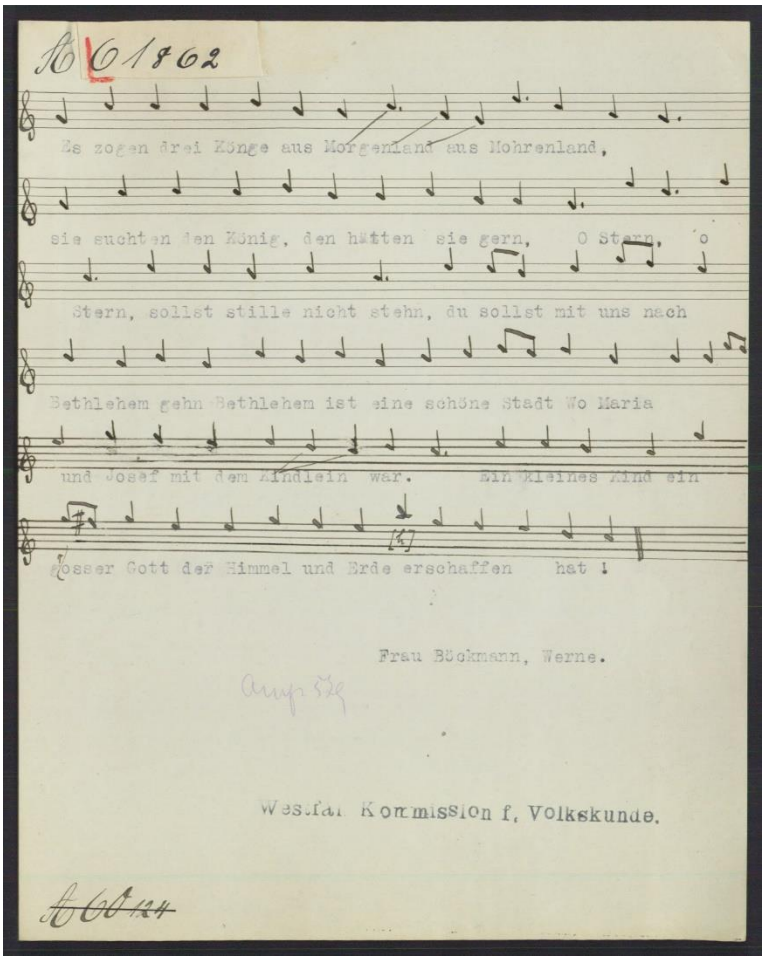
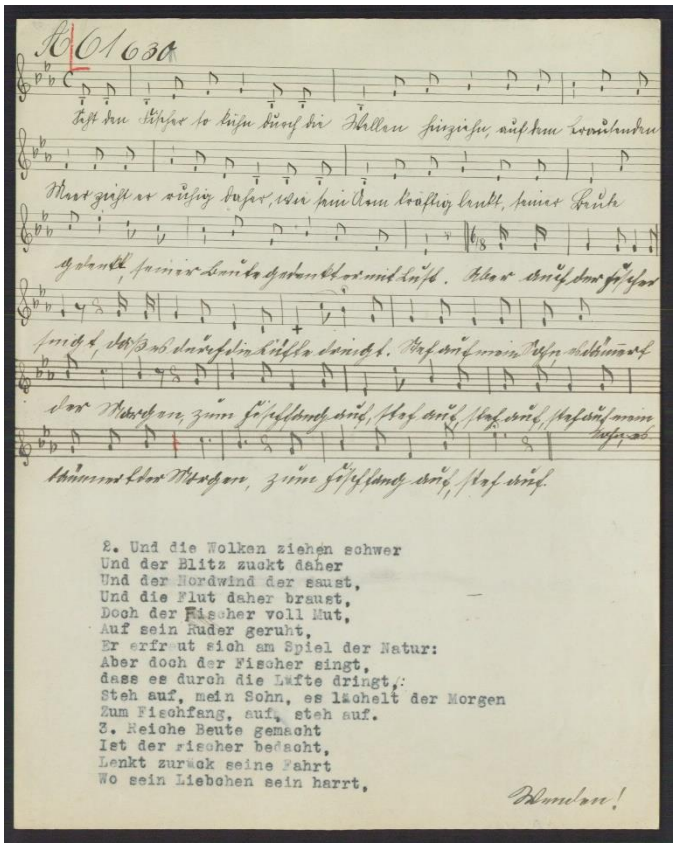
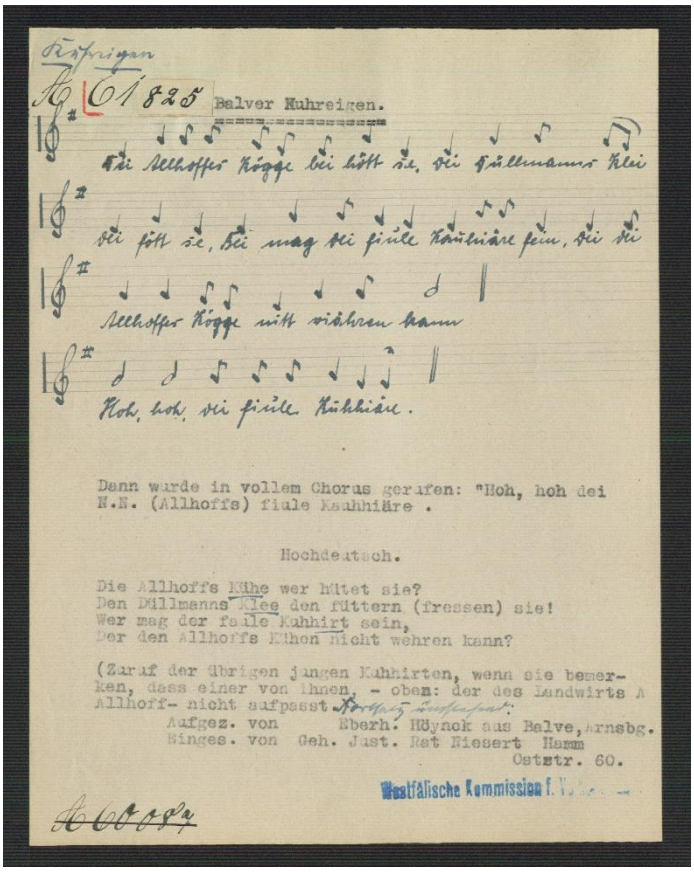


Abbildung 43: Beispiel Liedblatt Erkennungsrate 80% bei Photoscore



8. Und die Wolken ziehen schwer
Und der Blitz zuckt daher
Und der Nordwind der saust,
Und die Flut daher braust,
Doch der Fischer voll Mut,
Auf sein Ruder geruht,
Er erfreut sich am Spiel der Natur:
Aber doch der Fischer singt,
dass es durch die Lüfte dringt:
Steh auf, mein Sohn, es lächelt der Morgen
Zum Fischfang, auf, steh auf.
9. Reiche Beute gemacht
Ist der Fischer bedacht,
Lenkt zurück seine Fahrt
Wo sein Liebchen sein harrt,

Abbildung 44: Beispiel Liedblatt Erkennungsrate 13% bei Photoscore



Balver Kuhreigen.

Dann wurde in vollem Chorus gerufen: "Hoh, hoh dei N.N. (Allhoffs) fiale Kauhhiäre .

Hochdeutsch.

Die Allhoffs Kühe wer hütet sie?
Den Dillmanns Klee den füttern (fressen) sie!
Wer mag der faule Kuhhirt sein,
Der den Allhoffs Kühen nicht wehren kann?

(Zuruf der übrigen jungen Kuhhirten, wenn sie bemerken, dass einer von ihnen, - oben: der des Landwirts A Allhoff- nicht aufpasst.

Aufgez. von Oberh. Hüynck aus Balve, Arnsbg.
Einges. von Geh. Just. Rat Niessert Hamm
Oststr. 60.

Westfälische Kommission f. V...

Abbildung 45: Beispiel Liedblatt Erkennungsrate 0% bei Photoscore

Die Abbildungen zeigen außerdem, dass der Erfolg von OMR maßgeblich von der Notenkopfgröße und der allgemeinen Notendicke abhängt. Darüber hinaus wirkt sich Gleichmäßigkeit im Schriftbild positiv auf das Ergebnis aus. Eine detaillierte Analyse kann man mit Hilfe der Korpus-Tabelle und den entsprechenden Erkennungsraten unternehmen. Es handelt sich dabei um die Erkennungsraten für Noten, da diese am wichtigsten für die Erschließung sind, und um das Tool Photoscore, da dieses nachweislich die beste Leistung vollbrachte.

Tabelle 30: Test-Korpus mit Erkennungsraten

Signatur	Schriftbild	Notenkopf	Notenhals	Notenlinien	Fremdzeichen	Kontrast	Erkennungsrate
A59389	sauber	klein	lang	normal	nein	hoch	80%
A59394	sauber	klein	lang	normal	nein	hoch	60%
A59465	unsauber	klein	kurz	normal	nein	normal	43%
A59906	sauber	klein	kurz	normal	ja	hoch	48%
A60022	sauber	groß	lang	normal	nein	hoch	0%
A60051	unsauber	groß	kurz	normal	ja	hoch	58%
A61630	sauber	klein	kurz	normal	ja	gering	13%
A61815	sauber	groß	lang	normal	ja	hoch	43%
A61816	unsauber	groß	kurz	normal	nein	hoch	61%
A61818	unsauber	groß	lang	hell	nein	normal	0%
A61825	unsauber	groß	kurz	hell	ja	normal	0%
A61826	unsauber	groß	kurz	hell	nein	normal	0%
A61827	unsauber	groß	lang	hell	ja	normal	0%

A61833	unsauber	groß	lang	normal	nein	hoch	61%
A61858	unsauber	groß	lang	normal	ja	normal	0%
A61862	sauber	groß	lang	normal	ja	hoch	81%
A60019	sauber	groß	lang	normal	ja	hoch	52%
A60060	unsauber	groß	kurz	normal	ja	gering	0%
A61852	unsauber	groß	kurz	normal	ja	normal	67%
A61869	unsauber	groß	lang	normal	nein	hoch	70%

Man erkennt, dass die 7 Blätter, die eine 0%-Rate haben, also gar nicht als Notenblätter erkannt werden, mehrheitlich ein unsauberes Schriftbild und helle, kaum sichtbare Notenlinien haben. Die Beschaffenheit der Notenlinien ist quasi das Ausschlusskriterium. Jedes Liedblatt mit hellen Notenlinien (A61818 - A61827) wurde gar nicht erkannt. Dies ist bei den anderen Tools ähnlich. Bei den guten Ergebnissen sind die Kriterien heterogener verteilt. Alle Blätter über 60% haben gemein, dass sie einen hohen Kontrast haben, Notenhälse eher lang sind und die Notenlinien sichtbar. Es ist auch hilfreich, dass das Schriftbild sauber ist und keine Fremdzeichen vorhanden sind. Der Einfluss des Kontrasts weist auf eine mögliche Verbesserung der Ergebnisse über Bildbearbeitung hin.

Es konnte jedoch auch in Kapitel 3 gezeigt werden, dass selbst wenn man die sehr schlechten Blätter aus der Statistik „herausnimmt“, die Erkennungsrate mit ca. 50% noch immer weit unter Zielwert liegt. Eine vorhergehende Unterteilung von Liedblättern in Blätter, die für OMR geeignet sind und solche die ungeeignet sind, kann prinzipiell, nach den genannten Kriterien, vorgenommen werden. Der Einsatz von OMR-Tools ist jedoch, generell, in beiden Gruppen zweifelhaft.

Darüber hinaus sei anzumerken, dass im Rahmen dieses Projekts nicht in Erfahrung gebracht werden konnte, wie heterogen der große Bereich der Liedblätter außerhalb des hier genutzten Testkorpus ist. Es ist einerseits möglich, dass ein Großteil der 20 000 Blät-

ter bessere Ergebnisse erzielen würde. Andererseits ist davon auszugehen, dass die restliche Liedblattsammlung hinsichtlich der Blattbeschaffenheit ähnlich unterschiedlich ist, wie der in dieser Arbeit behandelte Testkorpus. Auch wurde gezeigt, dass nicht erkennbare Notenlinien bei diversen Scans zu einer Notenerkennungsrate von 0% geführt haben. Der Gesamtanteil der davon betroffenen Blätter an der ganzen Liedblattsammlung konnte im Rahmen dieser Arbeit nicht ausgemacht werden.

Angesichts der in Kapitel 3 beschriebenen Erkennungsraten kommt dieses Projekt zu dem Ergebnis, dass ein Crowdsourcing-Ansatz mit dem Ziel der Digitalisierung als vielversprechender anzusehen ist, als die Verwendung eines OMR-Programms. Grund dafür ist, dass die händische Verbesserung der OMR-Ausgabe genauso viel oder mehr Zeit in Anspruch nehmen würde, als die Musik von vornherein händisch zu digitalisieren.

Mehr zu den Problemen des Datenbestands und wie man damit umgehen sollte, kann man in Kapitel 4 des OCR-Berichts nachlesen. Holley (2009) empfiehlt als mögliche Lösung, die semi-automatische Erschließung, die im vorliegenden Projekt über ein Crowdsourcing-Tool realisiert werden soll.

5 Ausblick

Die in diesem Dokument beschriebene Evaluation dient der Fragestellung, ob Optical Music Recognition sinnvoll in einen Workflow eingebaut werden kann, um die vorliegende Liedblattsammlung effizient und möglichst maschinell zu digitalisieren. Die Ergebnisse, vor allem bezogen auf die Erkennungsraten, können somit keinesfalls als allgemeingültig für die jeweiligen OMR-Tools angesehen werden. Darüber hinaus wurde explizit auf die Anwendung diverser Bildbearbeitungssoftware auf die Scans verzichtet, selbst wenn dadurch eventuell ein besseres Ergebnis aufgrund optimierter Kontrast- oder Farbeinstellungen hätte erzielt werden können. Hinsichtlich des gesamten Workflows erscheint es problematisch, dass bei einer Sammlung von 20 000 Liedblättern eine manuelle Bildvorbearbeitung die Effizienz des Digitalisierungsvorgangs steigern würde. Hinzu kommt, dass das Ergebnis selbst bei dem besten Tool Photoscore mit 35% Notenerkennung so schlecht ist, dass OMR bei einer Erkennungssteigerung von 30% durch Bildvorbearbeitung für eine maschinelle Digitalisierung der Liedblattsammlung nach wie vor keine Rolle spielt.

Zur Verbesserung der Evaluation von OMR-Programmen könnte man jedoch noch diese Bildbearbeitungsschritte untersuchen. Auch wäre es interessant Open-Source Programme wie Audiveris miteinzubeziehen. Eine Vergrößerung des Korpus und damit des Stichprobenumfangs würde die Aussagekraft der statistischen Analysen erhöhen. Bellini, Bruno und Nesi (2007) haben ein Rechen-Modell entwickelt, bei dem man durch konkrete Zeitangaben für die Verbesserung von Fehlern die Dauer der Erschließung der Liedblattsammlung ohne Tool-Unterstützung exakt bestimmen kann. Eine Studie zur Erhebung der Zeitparameter wäre dann vonnöten. Mit dem Ergebnis kann man jedoch dann exakte zeitliche Aussage machen und die Probleme der Erschließung verdeutlichen. Ein Vergleich mit dem endgültigen Tool kann den Erfolg oder Misserfolg des vorliegenden Projektziels mit den Daten evaluieren.

Literaturverzeichnis

Alexandov, V. (2003). Error Evaluation and Applicability of OCR Systems. In: *International Conference on Computer Systems and Technologies - CompSysTech'2003.* New York: ACM Press.

Bainbridge, D. & Bell, T. (2001). The Challenge of Optical Music Recognition. *Computers and the Humanities,* 35, 95-121.

Bellini, P., Bruno, I. & Nesi, P. (2007). Assessing Optical Music Recognition Tools. *Computer Music Journal*, 31(1), 68-93.

Bortz, J. (2005). *Statistik für Human- und Sozialwissenschaftler.* (6. Auflage). Berlin: Springer-Verlag.

Carrasco, R. C. (2014). An open-source OCR evaluation tool. In: *DATeCH 2014*. New York: ACM Press.

Holley, R. (2009). How Good Can It Get? Analysing and Improving OCR Accuracy in Large Scale Historic Newspaper Digitisation Programs. *D – Lib Magazine*, 15(3/4). Retrieved from http://www.dlib.org/dlib/march09/holley/03holley.html

Kanungo, T., Marton, G. & Bulbul, O. (1998). *Paired Model Evaluation of OCR Algorithms*. Retrieved from http://www.dtic.mil/dtic/tr/fulltext/u2/a458678.pdf

Leonhart, R. (2010). *Datenanalyse mit SPSS*. Göttingen: Hogrefe Verlag.

Leonhart, R. (2013). *Lehrbuch Statistik: Einstieg und Vertiefung*. Bern: Verlag Hans Huber.

Lüpsen, H. (2015). *Varianzanalysen - Prüfen der Voraussetzungen und nichtparametrische Methoden sowie praktische Anwendungen mit R und SPSS*. Retrieved from http://www.uni-koeln.de/~a0032/statistik/texte/nonpar-anova.pdf

Rebelo, A., Fujinaga, I., Paszkiewicz, F., Marcal, A.R.S., Guedes, C. & Cardoso, J.S. (2012). Optical Music Recognition - State-of-the-Art and Open Issues. *International Journal of Multimedia Information Retrieval,* 1(3), 173-190.

Anhang

Evaluation

- Test-Korpus-Jpegs (alle Liedblätter als jpegs)
- Test-Korpus-Pdfs (alle Liedblätter als PDFs)

SPSS Daten und Auswertung

- Ausgangstabelle (als SPSS-Tabelle)
- Deskriptive Statistik (als SPSS-Viewer-Dateien)
- Inferenzstatistik (als SPSS-Viewer-Dateien)