



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 OCR-Tools mit der Hoerburger Liedblattsammlung

(vorläufige Fassung)

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

Inhalt

1	Einleitung	8
2	Methodisches Vorgehen.....	9
2.1	Testkorpus	9
2.2	Textzonen	12
2.3	„Grounded Truth“ – Dateien	13
2.4	OCR-Tools.....	14
2.5	Evaluation	15
2.5.1	Vorgehen	15
2.5.2	Erhobene Variablen.....	17
3	Ergebnisse.....	19
3.1	Testkorpus	20
3.2	ABBY – Deskriptive Statistik.....	20
3.3	Omnipage Professional – Deskriptive Statistik.....	29
3.4	Adobe Acrobat X Pro – Deskriptive Statistik.....	37
3.5	Vergleich – Deskriptive Statistik und Inferenzstatistik	45
3.5.1	Hypothesenformulierung	45
3.5.2	Signifikanztest – Statistisches Vorgehen	46
3.5.3	Voraussetzungen	47
3.5.4	Correct In Percent – Beispielhafter Ablauf	48
3.5.4.1	<i>Signifikanztests – Correct In Percent.....</i>	<i>48</i>
3.5.4.2	<i>Boxplot-Grafik – Correct In Percent</i>	<i>54</i>
3.5.5	Precision.....	56
3.5.5.1	<i>Signifikanztests – Precision.....</i>	<i>56</i>
3.5.5.2	<i>Boxplots-Grafik – Precision.....</i>	<i>57</i>
3.5.6	Spurious In Percent.....	58
3.5.6.1	<i>Signifikanztests – Spurious In Percent.....</i>	<i>58</i>
3.5.6.2	<i>Boxplots-Grafik – Spurious In Percent.....</i>	<i>60</i>
3.5.7	Confused In Percent.....	60
3.5.7.1	<i>Signifikanztests – Confused In Percent.....</i>	<i>60</i>
3.5.7.2	<i>Boxplots-Grafik – Confused In Percent.....</i>	<i>62</i>
3.5.8	Lost In Percent.....	63
3.5.8.1	<i>Signifikanztests – Lost In Percent.....</i>	<i>63</i>
3.5.8.2	<i>Boxplots-Grafik – Lost In Percent.....</i>	<i>64</i>
3.5.9	Character Error Rate	65
3.5.9.1	<i>Signifikanztests – CER.....</i>	<i>65</i>
3.5.9.2	<i>Boxplots-Grafik – CER.....</i>	<i>66</i>
3.5.10	Nachbemerkung – Frakturschrift	67

3.5.11	Weitere vergleichende Visualisierungen.....	67
3.5.11.1	<i>Scatterplots</i>	67
3.5.11.2	<i>Gestapelte Balkendiagramme</i>	74
4	Diskussion	77

Abbildungen

Abbildung 1: ubr16444_0170.....	10
Abbildung 2: ubr16444_0101 – Frakturschrift.....	11
Abbildung 3: ubr16444_0525 – sehr helle, kontrastarme Schrift	12
Abbildung 4: Begrenzte Textzone auf Liedblatt	13
Abbildung 5: Grounded Truth Text für Datei ubr16444_0121	13
Abbildung 6: ABBY-txt-Output für Datei ubr16444_0121.....	15
Abbildung 7: HTML-Output von ocrevalUation	16
Abbildung 8: Grafische Visualisierung von ocrevalUation	16
Abbildung 9: Histogramm – ABBY Correct In Percent.....	21
Abbildung 10: Histogramm – ABBY Precision	22
Abbildung 11: ABBY – Spurious in Percent.....	23
Abbildung 12: Histogramm – ABBY Confused In Percent.....	24
Abbildung 13: Histogramm – ABBY Lost In Percent.....	25
Abbildung 14: Histogramm – ABBY CER.....	26
Abbildung 15: Kreisdiagramm – ABBY Complete.....	27
Abbildung 16: Kreisdiagramm – ABBY Grounded Truth.....	28
Abbildung 17: Kreisdiagramm – ABBY OCR-Output.....	28
Abbildung 18: Histogramm – Omnipage Correct In Percent.....	30
Abbildung 19: Histogramm – Omnipage Precision.....	31
Abbildung 20: Histogramm – Omnipage Spurious In Percent.....	32
Abbildung 21: Histogramm – Omnipage Confused In Percent	33
Abbildung 22: Histogramm – Omnipage Lost In Percent.....	34
Abbildung 23: Histogramm – Omnipage CER.....	35
Abbildung 24: Kreisdiagramm – Omnipage Complete	36
Abbildung 25: Kreisdiagramm – Omnipage Grounded Truth	36
Abbildung 26: Kreisdiagramm – Omnipage OCR-Output	37
Abbildung 27: Histogramm – Acrobat Correct In Percent	38
Abbildung 28: Histogramm – Acrobat Precision	39
Abbildung 29: Histogramm – Acrobat Spurious In Percent	40
Abbildung 30: Histogramm – Acrobat Confused In Percent.....	41
Abbildung 31: Histogramm – Acrobat Lost In Percent	42
Abbildung 32: Histogramm – Acrobat CER.....	43
Abbildung 33: Kreisdiagramm – Acrobat Complete.....	44
Abbildung 34: Kreisdiagramm – Acrobat Grounded Truth	44
Abbildung 35: Kreisdiagramm – Acrobat OCR-Output.....	45

Abbildung 36: Boxplot-Grafik – Correct In Percent	55
Abbildung 37: Boxplot Beispiel.....	55
Abbildung 38: Boxplot-Grafik – Precision	58
Abbildung 39: Spurious In Percent	60
Abbildung 40: Boxplot-Grafik – Confused In Percent.....	62
Abbildung 41: Boxplot-Grafik – Lost In Percent	64
Abbildung 42: Boxplot-Grafik – CER.....	66
Abbildung 43: Scatterplot Interpretation	68
Abbildung 44: Scatterplot – Correct In Percent ABBY-Omnipage	69
Abbildung 45: Scatterplot – Correct In Percent ABBY/Acrobat.....	69
Abbildung 46: Scatterplot – Correct In Percent Omnipage/Acrobat.....	70
Abbildung 47: Scatterplot – Lost In Percent ABBY-Omnipage.....	71
Abbildung 48: Scatterplot – Lost In Percent ABBY/Acrobat.....	71
Abbildung 49: Scatterplot – Lost In Percent Omnipage-Acrobat	72
Abbildung 50: Scatterplot – CER ABBY/Omnipage.....	73
Abbildung 51: Scatterplot – CER ABBY/Acrobat.....	73
Abbildung 52: Scatterplot – CER Omnipage/Acrobat	74
Abbildung 53: Gestapeltes Balkendiagramm Zeichentypen – Grounded Truth.....	75
Abbildung 54: Gestapeltes Balkendiagramm Zeichentypen – OCR-Output	76
Abbildung 55: Gestapeltes Balkendiagramm Zeichentypen – Complete	77

Tabellen

Tabelle 1: Testkorpus	9
Tabelle 2: Deskriptive Statistik – Total	20
Tabelle 3: Deskriptive Statistik – ABBY.....	21
Tabelle 4: Deskriptive Statistik – Omnipage	29
Tabelle 5: Deskriptive Statistik – Acrobat X Pro.....	38
Tabelle 6: Deskriptive Statistik – Correct In Percent.....	48
Tabelle 7: Mauchly-Test auf Sphärizität – Correct In Percent.....	49
Tabelle 8: Varianzanalyse mit Messwiederholung – Correct In Percent.....	49
Tabelle 9: Zusammenhangstyp – Correct In Percent.....	50
Tabelle 10: Deskriptive Statistik – Rangtransformation Correct In Percent.....	50
Tabelle 11: Friedman-Test – Correct In Percent.....	51
Tabelle 12: Paarweise Signifikanztests – Correct In Percent.....	51
Tabelle 13: Deskriptive Statistik – Rangtransformation Paarweise I	52
Tabelle 14: Friedman-Test – Paarweise I	52
Tabelle 15: Deskriptive Statistik – Rangtransformation Paarweise II	52
Tabelle 16: Friedman-Test – Paarweise II	52
Tabelle 17: Deskriptive Statistik – Rangtransformation Paarweise III	53
Tabelle 18: Friedman-Test – Paarweise III	53
Tabelle 19: Signifikanztests Zusammenfassung – Correct In Percent	53
Tabelle 20: Paarweise Signifikanztests Zusammenfassung – Correct In Percent.....	54
Tabelle 21: Deskriptive Statistik – Precision.....	56
Tabelle 22: Signifikanztests Zusammenfassung – Correct In Percent	57
Tabelle 23: Paarweise Signifikanztests Zusammenfassung – Precision	57
Tabelle 24: Deskriptive Statistik – Spurious In Percent.....	58
Tabelle 25: Signifikanztests Zusammenfassung – Spurious In Percent	59
Tabelle 26: Paarweise Signifikanztests Zusammenfassung – Spurious In Percent.....	59
Tabelle 27: Deskriptive Statistik – Confused In Percent	61
Tabelle 28: Signifikanztests Zusammenfassung – Confused In Percent.....	61
Tabelle 29: Paarweise Signifikanztests Zusammenfassung – Confused In Percent.....	61
Tabelle 30: Deskriptive Statistik – Lost In Percent	63
Tabelle 31: Signifikanztests Zusammenfassung – Lost In Percent	63
Tabelle 32: Paarweise Signifikanztests Zusammenfassung – Lost In Percent.....	63
Tabelle 33: Deskriptive Statistik – CER	65
Tabelle 34: Signifikanztests Zusammenfassung – CER.....	65

Tabelle 35: Paarweise Signifikanztests Zusammenfassung – CER.....	65
---	----

1 Einleitung

Der Universität Regensburg steht eine Liedblattsammlung, im Umfang von ca. 20 000 Liedblättern unterschiedlicher deutschsprachiger Volksmusik, zur Verfügung. Diese Liedblätter bestehen aus Notenzeilen und Texten mit unterschiedlichen Informationsgehalt (Liedtexte, Archivstempel usw.). Um die Liedblattsammlung für Musikwissenschaftler und Ethnologen zugänglich zu machen, ist geplant die Sammlung komplett, mit allen Informationen (vor allem Musik und Text), die ein Liedblatt bietet, zu digitalisieren. Dafür soll ein effizienter Digitalisierungs-Workflow eingerichtet werden bzw. der momentane erweitert werden.

Beim bisherigen Workflow wird auch OCR, Optical Character Recognition eingesetzt. Unter Optical Character Recognition (optische Zeichenerkennung oder Texterkennung) versteht man alle Verfahren der maschinellen und automatischen Texterkennung in Bildern. Dabei kann es sich um gedruckten oder handschriftlichen Text handeln. Es gibt unterschiedliche kommerzielle und nicht-kommerzielle Tools und Programme, mit denen man OCR durchführen kann. Das OCR-Ergebnis kann meist als txt-Datei oder in einem xml-Format mit zusätzlichen Meta-Informationen ausgegeben werden.

Zu diesem Zweck wurde im Rahmen dieses Projekts die Frage untersucht inwiefern man die textuelle Erschließung mittels OCR-Tools durchführen kann. Ob dies also überhaupt zielführend ist und welches Tool in diesem Fall am besten geeignet ist. Auf Basis einer vergleichenden Evaluationsstudie werden Schlussfolgerungen gezogen wie mit der textuellen Erschließung weiter zu verfahren ist. Eine äquivalente Studie, jedoch mit einem deutlich begrenzten Test-Korpus, fand auch für die automatische Musikererkennung statt.

Der Arbeit liegt ein Anhang in digitaler Form bei. Dieser enthält alle Dateien für die Durchführung der Studie, statistische Analysen und Grafiken (meist in SPSS-Format) sowie die in dieser Studie genutzten Programme.

Mehr Informationen zum Projekt findet man auch in der dazugehörigen Software Requirements Specification (SRS).

Von einer erschöpfenden Literaturlaufbereitung wurde (zum Zeitpunkt 7.8.2015, aus zeitlichen Gründen) abgesehen. Die für die Studie wichtige Literatur wird im weiteren Text referenziert.

2 Methodisches Vorgehen

In diesem Abschnitt wird das experimentelle Setup für die OCR-Evaluation beschrieben. Gemäß Kanungo, Marton und Bulbul (1999) handelt es sich bei dem hier vorliegenden Vorgehen um eine Blackbox-Evaluation. Dabei wird ein OCR-System als unzerteilbare Einheit betrachtet und das System nur in Hinblick auf das Endergebnis betrachtet (Bei „Whitebox“-Evaluationen werden einzelne Systemkomponenten untersucht). Die Studie ist komparativ, da unterschiedliche Tools hinsichtlich ihrer Performanz verglichen werden. Nach Kanungo et al. (1999) gliedert sich der Ablauf einer solchen Evaluation in folgende Teile:

- Erstellung des Korpus
- Abgrenzung der Textzone oder Textzonen, die untersucht werden
- Erstellung des korrekten Textes für diese Textzone oder Textzonen
- Ausführung des OCR eines jeden Systems
- Vergleich des OCR-Ergebnisses mit dem korrekten Text

Die einzelnen Schritte werden nun im Folgenden näher beschrieben

2.1 Testkorpus

Der Testkorpus besteht aus 102 jpeg-Dateien aus der vom Universitätsarchiv zur Verfügung gestellten Scans. Die jpeg-Dateien sind demnach schon über den Workflow des Universitätsarchivs für das OCR geringfügig vorbereitet (z.B. Ausrichtung). Es handelt sich dabei um abschnittsweise gezogene Sequenzen aus dem Originalkorpus. Die Sequenzen wurden so gewählt, dass die große Heterogenität des Datenbestands repräsentiert ist und möglichst viele Eigenheiten vertreten sind. Leere Blätter wurden aus der Sammlung entfernt. Bei Liedblättern, die Text über mehrere Seiten beinhalten, wurde jede Seite als einzelne Einheit betrachtet. Folgende Tabelle zeigt alle vier Sequenzen auf:

Tabelle 1: Testkorpus

Sequenz	Dateiname von	Dateiname bis	Signatur von	Signatur bis	Besonderheit	Zahl der Seiten
1	Ubr16444_0087	Ubr16444_0135	A59271	A59302	Frakturschrift, viel aufgeklebt	25

2	Ubr16444_0291	Ubr16444_0365	A59422	A59512	Druckschrift, zwischen Zeilen und übereinander	25
3	Ubr16444_0481	Ubr16444_0529	A60021	A60044	Druckschrift, zwischen Zeilen und übereinander	27
4	Ubr16444_0691	Ubr16444_0728	A61848	A61866	Druckschrift, sehr schwach, viele Sonderzeichen	25

Der Testkorpus ist im digitalen Anhang beigefügt. Hier seien einige Beispiele präsentiert, die die unterschiedlichen Manifestationen von Liedblättern darstellen:

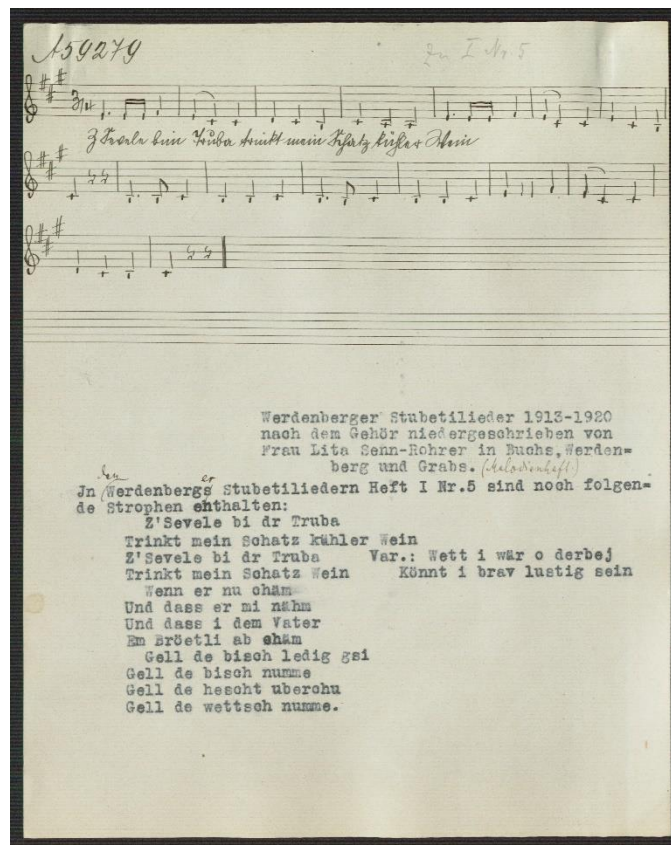


Abbildung 1: ubr16444_0170

159280

Koti Chriasi han i gern

[Dollmetschung mit
Lita 1. St.]

Koti Chriasi han i gern

1. Sie: Koti Chriasi han i gern Koti Chriasi
Di Schwarga no vil lieber Di Schwarga
D' Stadiner Buaba han i gern Buaba
Und Bucher no viel lieber.

2. Er: Meitali, wenn hürta witt Meitali
Hört en Oberrieter
Und wenn en rechta Chroepfi isch Meitali
So mues er nid go chiega. Meitali

3. Sie: Hunderttufig Opfellschneig Meitali
Das git en groesja Hufte. Meitali
Und wenn der Buab zum Meitali got Meitali
So tar er numma juffa. Meitali

4. Er: Miner Mueter Kaffimägli Meitali
Rumpfet um und numme Meitali
Schäpali, wend mi du nid witt Meitali
So will i di o numme. Meitali

5. Sie: Ali Wägali fingen schide Meitali
Bis am Sametig jobet Meitali
Ali Wäbali hetten mi gern Meitali
D wie bin i ploget. Meitali

6. Er: Lustig wemmo ledig isch Meitali
Und lustig vor de Lita Meitali
Und wers eim nit verträga mag Meitali
Der has eim jo verbüta. Meitali

Verdenberger Stubetlieder.
Lita Sonn-Rohrer
1871

Abbildung 2: ubr16444_0101 – Frakturschrift

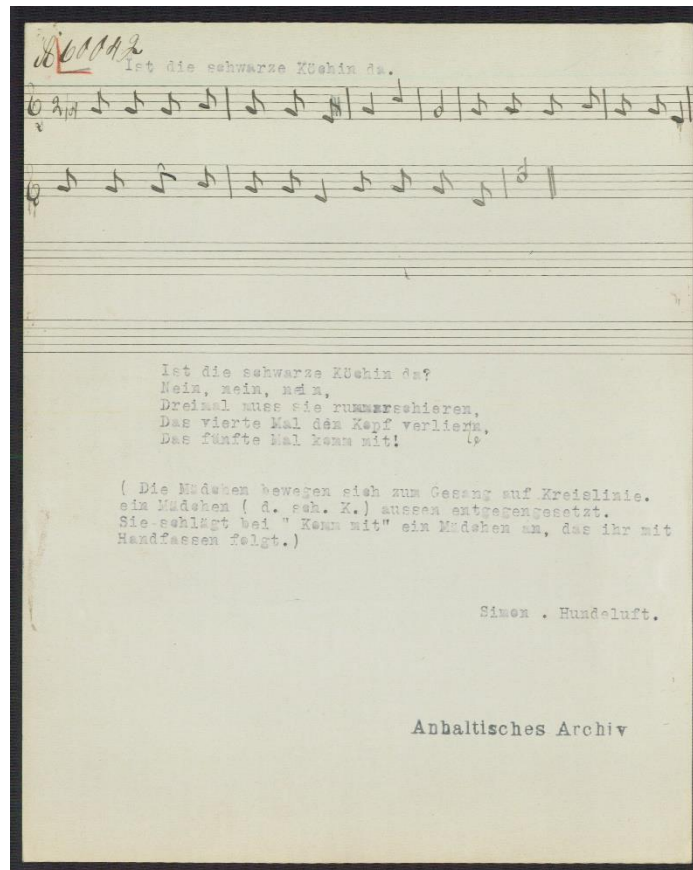


Abbildung 3: ubr16444_0525 – sehr helle, kontrastarme Schrift

2.2 Textzonen

Um die Blätter vereinfacht zu vergleichen, hat man sich bei der Evaluation eine Textzone pro Blatt ausgewählt. Die Textzone, die ausgewählt wurde, ist die Textzone gleich unter den Notenlinien. Da diese den meisten Text in Druckschrift enthält und bei jedem Blatt relativ konsistent vorkommt, ist diese am besten zum Vergleich geeignet. Sie besteht meistens aus Liedtexten, Hinweisen zur Durchführung oder anderen Beschreibungen. Handschrift, Stempel und andere Besonderheiten wurden bewusst aus der Evaluation ausgeschlossen um den Vergleich zu vereinfachen, aber auch um in dieser ersten Studie die Leistung, in Bezug auf den größten und wichtigsten Bereich, zu untersuchen.

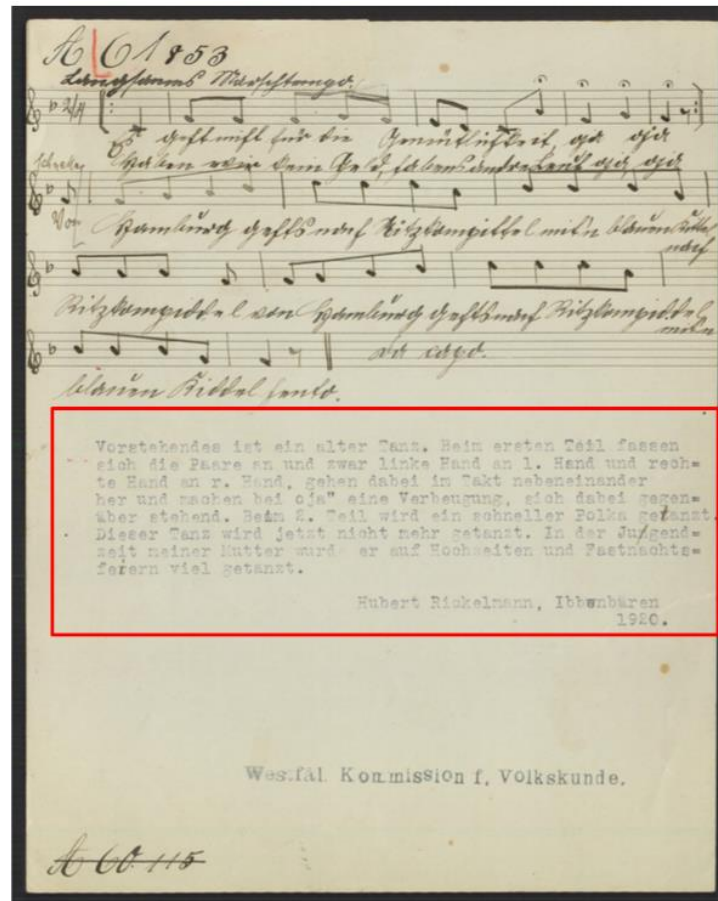


Abbildung 4: Begrenzte Textzone auf Liedblatt

Selten befand sich Liedtext zwischen den Notenzeilen. Für diesen Fall wurde der OCR-Output auch so formatiert, dass er äquivalent vergleichbar ist mit dem dazu erstellten Grounded-Truth-Text.

2.3 „Grounded Truth“ – Dateien

Als Grounded Truth bezeichnet man im OCR eine Dateiensammlung, die vollständig korrekten Text für eine Bilddatei enthält. In der hier vorliegenden Studie wurden für alle 102 jpeg-Dateien für die in Abschnitt 2.2 beschriebenen Textzonen händisch die korrekten Texte geschrieben und als txt-Dateien abgespeichert.

```

1 Der Ueli mit em Täsli
2 Der Ueli mit em Täsli
3 Nimm e du, i will ne nid.
4 I bi e Wili bei em gsi.
5 Het gmeint, i sött si Dieneri si;
6 Nimme nimme du, i will ne nid

```

Abbildung 5: Grounded Truth Text für Datei ubr16444_0121

2.4 OCR-Tools

Ziel der Evaluation ist nicht der direkte Vergleich der Tools, sondern überhaupt zu untersuchen inwiefern eine maschinelle Erschließung des Datenbestands mit Hilfe von OCR-Tools sinnvoll ist. Als OCR-Tools für den Vergleich hat man ABBY Fine Reader¹ in Form des Abby Servers, der dem Universitätsarchiv zur Verfügung steht, Omnipage² Professional und Adobe Acrobat X Pro³ hergenommen. Die Auswahl auf diese drei Tools ist vor allen Dingen von einer der Evaluation vorangehenden Web-Recherche beeinflusst. Unterschiedliche Quellen im Netz empfehlen diese drei Tools oder geben Sie als die qualitativ besten für OCR an (Top Ten Reviews, 2015; Fitzpatrick, 2010). Über diese Web-Recherche und eine Literatur-Recherche wurden unterschiedliche OCR-Tools gesammelt und ihre Funktionen und Besonderheiten tabellarisch zusammengefasst um basierend auf diesen Informationen eine annehmbare Auswahl zu treffen. Tools die jedoch Nachteile für einen effizienten Digitalisierungsprozess bieten (z.B. fehlendes Batchverfahren) wurden durch diese Analyse eliminiert (z.B. Google Docs⁴). Auch kommerzielle Tools, deren Testversionen nicht genügend Funktionen für eine Evaluation anbieten, wurden aus der Auswahl entfernt (z.B. Readiris⁵). Entscheidend für die Auswahl waren Empfehlungen aus der Web-Recherche und auch die Vertretung in wissenschaftlicher Forschung.

Sowohl Omnipage (Kanungo et al., 1999; Mello & Lins, 2012) als auch ABBY (Holley, 2009; Karaoglu, van Gerner & Gevers, 2012) werden in wissenschaftlichen Studien verwendet und weisen sehr gute Erkennungsraten auf (>90%). Alle drei Tools sind kommerzielle Produkte. Im Fall von ABBY hat man den Output des ABBY Servers hergenommen. Für Omnipage und Adobe Acrobat X Pro wurden Testversionen verwendet, die jedoch die vollständige Funktionalität für den Testzeitraum geboten haben.

Von ABBY und Omnipage wurde der OCR-Output in Form einer txt-Datei erstellt. XML-Output mit deutlich mehr Informationsgehalt wie z.B. Positionierung von Textblöcken ist bei beiden Tools möglich. Für Acrobat X Pro wird eine PDF erstellt. Von dieser

¹ <http://www.abbyy.de/>

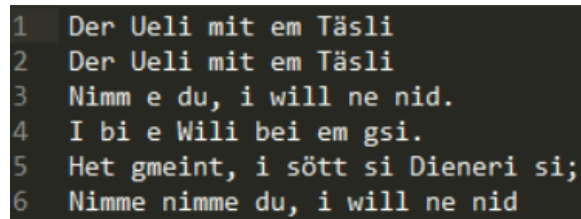
² <http://www.nuance.de/for-individuals/by-product/omnipage/index.htm>

³ <https://helpx.adobe.com/de/acrobat/kb/acrobat-downloads.html>

⁴ <https://www.google.de/intl/de/docs/about/>

⁵ <http://www.irislink.com/c1-3089-48/Readiris-15---OCR-Software--Kein-Eintippen--Kein-Papier--Nur-smarte-Dokumente.aspx>

wurde der Inhalt händisch in eine txt-Datei kopiert. Jede txt-Datei für jedes Tool wurde danach noch händisch auf den zu untersuchenden Textblock formatiert, indem umliegende Noise und fremde Textblöcke entfernt wurden, so dass die txt-Dateien mit den Grounded-Truth-Dateien vergleichbar sind.



```
1 Der Ueli mit em Täsli
2 Der Ueli mit em Täsli
3 Nimm e du, i will ne nid.
4 I bi e Wili bei em gsi.
5 Het gmeint, i sött si Dieneri si;
6 Nimme nimme du, i will ne nid
```

Abbildung 6: ABBY-txt-Output für Datei ubr16444_0121

Die Grounded-Truth-Dateien sind im Anhang enthalten.

2.5 Evaluation

2.5.1 Vorgehen

Zur Evaluation wurde für jedes Tool der txt-Output mit der jeweiligen Grounded-Truth-Datei verglichen. Als Hilfe für diesen Prozess hat man das Tool `ocrevalUAtion`⁶ verwendet (Carassco, 2014). Das Tool vergleicht zwei Text-Dateien, einen Grounded-Truth-Text und einen OCR-Output und generiert einen html-Report mit einigen OCR-Evaluations-Statistiken, sowie eine grafische Visualisierung der Unterschiede beider Texte. Vor Verwendung wurde das Tool stichpunktartig auf korrekte Arbeitsweise überprüft. Das Tool ist eine Sammlung von java-Klassen die jedoch auch über eine Konsole verwendet werden können (über die `jar ocrevalUAtion-1.3.0-jar-with-dependencies.jar`). Folgender Befehl führt für die beiden Dateien `ubr16444_0299.txt` und `ubr16444_0299.2.txt` (Grounded Truth) einen Vergleich durch und speichert das Ergebnis in den Ordner „zwischenordner“:

```
java -cp ocrevalUAtion-1.3.0-jar-with-dependencies.jar eu.digitisation.Main -gt ubr16444_0299.2.txt -ocr ubr16444_0299.txt -d zwischenordner
```

⁶ <https://github.com/impactcentre/ocrevalUAtion>

Das Tool wurde für jedes Dateienpaar ausgeführt und die Ergebnisse in Form der html-Oupputs gesichert. Diese befinden sich im digitalen Anhang.

General results

CER	36.31
WER	73.68
WER (order independent)	60.53

Difference spotting

ubr16444_0121.2.txt	ubr16444_0121.txt
Der Ueli mit em Täsl Der Ueli mit em Täsl Nimm e du, i will ne nid. I bi e Wili bei em gsi. Het gmeint, i sött si Dieneri si; Nimm nimm du, i will ne nid	5 ft iteli mit rat @Äsl TTn-5<& Ser Ueli mit em \$Mi SRimm e bu, i mH ne nib. 3 bi e SSili bei em gfi jet gmeint, i fött fi SDieneri fi; Stimme nimm bu, i mifl ne nib

Error rate per character and type

Character	Hex code	Total	Spurious	Confused	Lost	Error rate
0	30	37	1	0	0	2.70
\$	24	0	1	0	0	Infinity
ä	26	0	1	0	0	Infinity
ˆ	27	0	1	0	0	Infinity
2c	3	0	0	0	0	0.00
ˆ	28	0	1	0	0	Infinity
2e	2	1	0	1	1	100.00
5	35	0	1	0	0	Infinity
ˆ	3b	1	0	0	0	0.00
ˆ	3c	0	1	0	0	Infinity
D	44	3	0	2	0	66.67
H	48	1	0	1	0	100.00
I	49	1	0	1	0	100.00
N	4e	2	0	2	0	100.00
R	52	0	1	0	0	Infinity
S	53	0	2	0	0	Infinity
T	54	2	1	2	0	150.00
U	55	2	0	1	0	50.00
W	57	1	0	1	0	100.00
b	62	2	0	0	0	0.00
d	64	4	0	4	0	100.00
e	65	18	0	3	0	16.67
f	66	0	1	0	0	Infinity
#	67	2	0	0	0	0.00

Abbildung 7: HTML-Output von ocrevalUAtion

Difference spotting

ubr16444_0121.2.txt	ubr16444_0121.txt
Der Ueli mit em Täsl Der Ueli mit em Täsl Nimm e du, i will ne nid. I bi e Wili bei em gsi. Het gmeint, i sött si Dieneri si; Nimm nimm du, i will ne nid	5 ft iteli mit rat @Äsl TTn-5<& Ser Ueli mit em \$Mi SRimm e bu, i mH ne nib. 3 bi e SSili bei em gfi jet gmeint, i fött fi SDieneri fi; Stimme nimm bu, i mifl ne nib

Abbildung 8: Grafische Visualisierung von ocrevalUAtion

Das Tool berechnet auch einige Metriken, die für die Evaluation unnötig waren, beispielsweise Korrektheit pro Buchstaben. Bezüglich der Vergleichs-Metriken hat man sich an Alexandrov (2003) orientiert. Einige Metriken werden jedoch nicht vom Tool berechnet bzw. müssen auf Basis der Daten noch berechnet werden. Dafür wurde ein eigenes Java-Programm htmlReader.java geschrieben. Das Tool erwartet einen Ordner von html-Dateien aus dem Tool ocrevalUAtion, übernimmt die vorhandenen Metriken und berechnet die fehlenden Metriken. Alle Informationen speichert es als xls-Datei. So

kann man für jedes Tool alle fehlenden Maße in einer Tabelle zusammenfassen. Das Programm befindet sich auch im Anhang. In Zeile 18 des Programms gibt man den Ursprungs-Ordner an und in Zeile 22 den Namen der Zielfeile.

Alle Metriken wurden in eine Tabelle übernommen und für gewisse Maße noch primitive Tabellenkalkulation durchgeführt.

2.5.2 Erhobene Variablen

Nachfolgend seien alle erhobenen Variablen und Maße (für eine Datei) näher erläutert. Es handelt sich dabei um eine Zusammenstellung verschiedener Metriken aus der Forschung (Alexandrov, 2003; Carrasco, 2014; Kanungo, Marton & Bulbol, 1998). Jedes Maß gilt für ein einzelnes Liedblatt:

Total:

Die totale Anzahl von Zeichen in der Grounded-Truth-Datei.

Spurious:

Die Zahl der überflüssigen Zeichen im OCR-Output (Noise), also Zeichen die in der Grounded-Truth-Datei nicht vorkommen. Es handelt sich um einen Fehler.

Confused:

Die Zahl der falsch erkannten Zeichen im OCR-Output. Es handelt sich um einen Fehler.

Lost:

Die Zahl der gar nicht erkannten Zeichen im OCR-Output. Es handelt sich um einen Fehler.

Correct:

Die Zahl der korrekt erkannten Zeichen der Zeichen aus dem Grounded Truth im OCR-Output.

Es gilt $\text{Total} - \text{Confused} - \text{Lost} = \text{Correct}$ für ein Liedblatt. Für den OCR-Output gilt, dass er gleich $\text{Total} + \text{Spurious}$ ist bzw. $\text{Correct} + \text{Confused} + \text{Spurious}$.

Spurious, Confused, Lost, Correct In Percent:

Der prozentuale Anteil des jeweiligen Maßes in Bezug auf den Grounded-Truth-Text. Es gilt also obige Formel auch für die Prozentangabe. Für die Statistik ist es bedeutungslos ob die Prozentwerte oder die Anteile kleiner 1 genommen werden. Die Prozentwerte sind intuitiv zugänglicher.

Spurious in Percent:

$$100 * \text{Spurious} / \text{Total}.$$

Es gilt je größer desto schlechter.

Confused in Percent:

$$100 * \text{Confused} / \text{Total}$$

Es gilt je größer desto schlechter.

Lost In Percent:

$$100 * \text{Lost} / \text{Total}$$

Es gilt je größer desto schlechter.

Correct In Percent:

$$100 * \text{Correct} / \text{Total}$$

Correct in Percent wird auch Accuracy oder Erkennungsrate genannt und wird am häufigsten als Metrik im OCR verwendet. Tatsächlich vernachlässigt Correct In Percent aber den Spurious-Wert. Dennoch zählt die Accuracy zu den Hauptmetriken und wird prominent in dieser Studie behandelt. Es gilt, je größer desto besser. Der Parameter verhält sich ähnlich zum Recall im IR.

Precision:

$$100 * \text{Correct} / (\text{Total} + \text{Spurious})$$

Der Anteil der korrekten, erkannten Zeichen im gesamten OCR-Output, im Gegensatz zu Correct in Percent, dass den Anteil an den Grounded-Truth-Zeichen berechnet. Ähnlich zur Precision im IR gibt es die Genauigkeit bei der Erkennung an und berechnet auch ein, wie viel Noise ausgegeben wird. Diese Metrik ist hilfreich wenn zwei Tools eine ähnliche Accuracy haben. Dasjenige mit der größeren Precision generiert weniger Noise.

CER (Character Error Rate) oder auch GER (Global Error Rate):

$$100 * (\text{Spurious} + \text{Confused} + \text{Lost}) / \text{Total}$$

Der prozentuale Anteil aller Fehler eines Systems gemessen an der Gesamtanzahl der Zeichen des Grounded Truth-Textes. Alle Fehler lassen sich durch die Summe von Spurious, Confused und Lost berechnen. Diese Variable ist demnach äquivalent zu *Spurious in Percent + Confused in Percent + Lost in Percent*. Die Variable kann demnach den Wert von 100% übersteigen, wenn ein Ergebnis einen besonders hohen Wert von Spurious aufweist. Sie gibt insgesamt den Fehleranteil in Bezug auf den

Original-Text an. Das Maß gilt als das Hauptmaß im OCR (Alexandrov, 2003). Es wirkt kritischer als Correct in Percent da hier noch die Spurious-Werte betrachtet werden.

Holley (2009) weist daraufhin, dass kein konkreter Konsens über die Interpretation der Maße besteht. Lediglich für die Accuracy gibt sie, auf Basis ihrer Meta-Studien, eine Empfehlung an:

<i>„Good OCR accuracy</i>	<i>= 98-99% accurate</i>	<i>(1-2% of OCR incorrect)</i>
<i>Average OCR accuracy</i>	<i>= 90-98% accurate</i>	<i>(2-10% of OCR incorrect)</i>
<i>Poor OCR accuracy</i>	<i>= below 90% accurate</i>	<i>(more than 10% of OCR incorrect)“</i>

Grund hierfür ist, dass eine Ausbesserung bei einer Erkennungsrate unter 90% aufwändiger wäre, als den Text komplett neu zu schreiben. Für CER wird eine analoge Annahme getroffen und bei einem Wert über 10-20% von schlechter Fehlerrate gesprochen. Andere Maße werden jedoch nicht angesprochen und deswegen hier individuell interpretiert.

Als Hauptmaße werden Correct In Percent (Accuracy), Precision und die Character Error Rate betrachtet. Tatsächlich gibt es noch mehr Maße wie z.B. die Word Error Rate oder Level of Reliability (Carrasco, 2014). Diese wurden jedoch entweder als nicht passend für den hier vorliegenden Anwendungsfall angenommen, sind lediglich ästhetischer Natur (Modified CER) oder redundant und wurden deswegen nicht weiter verarbeitet, wenn auch erhoben.

Alle Maße wurden tabellarisch erfasst und formatiert um sie in das Statistikprogramm SPSS⁷ zu übertragen, mit welchem die Auswertung durchgeführt wurde.

3 Ergebnisse

Im folgenden Abschnitt werden die Ergebnisse der Evaluation präsentiert. Als erstes wird der Testkorpus kurz statistisch erläutert. Danach werden zunächst die Leistungen jedes einzelnen Tools mit Hilfe von deskriptiver Statistik aufgezeigt. Die betrachteten Variablen sind dabei alle Fehlervariablen (Spurious, Lost, Confused) in Prozent, sowie

⁷ <http://www-01.ibm.com/software/de/analytics/spss/>

Correct In Percent, Precision und der CER. Die Verteilungen dieser Performanz-Variablen wird, gemäß der Empfehlung für Darstellungsformen für stetige Variablen von Leonhart (2013, S. 85), mittels Histogrammen und Kreisdiagrammen veranschaulicht. In einem weiteren Abschnitt werden die Tools mit Hilfe eines Signifikanztests für verbundene Stichproben verglichen und die Unterschiede über Visualisierungsmöglichkeiten der deskriptiven Statistik dargestellt.

3.1 Testkorpus

Der Testkorpus besteht aus 102 Liedblättern, also 102 Datensätzen. Zu den in Abschnitt 3.5 schon beschriebenen Variablen wurde noch die Liedblattnummer gespeichert. Alle Variablen zu einem Tool sind metrisch skaliert. Eine text-Datei hat im Schnitt 509 Zeichen. Es macht Sinn sich für alle Variablen mit den prozentualen Anteilen auseinanderzusetzen, da die Varianz sowie Minimum und Maximum auf eine starke Streuung hinweisen.

Tabelle 2: Deskriptive Statistik – Total

Deskriptive Statistiken						
	N	Minimum	Maximum	Mittelwert	Standardabweichung	Varianz
Total	102	57,00	1178,00	508,5196	245,46604	60253,579
Gültige Anzahl (listenweise)	102					

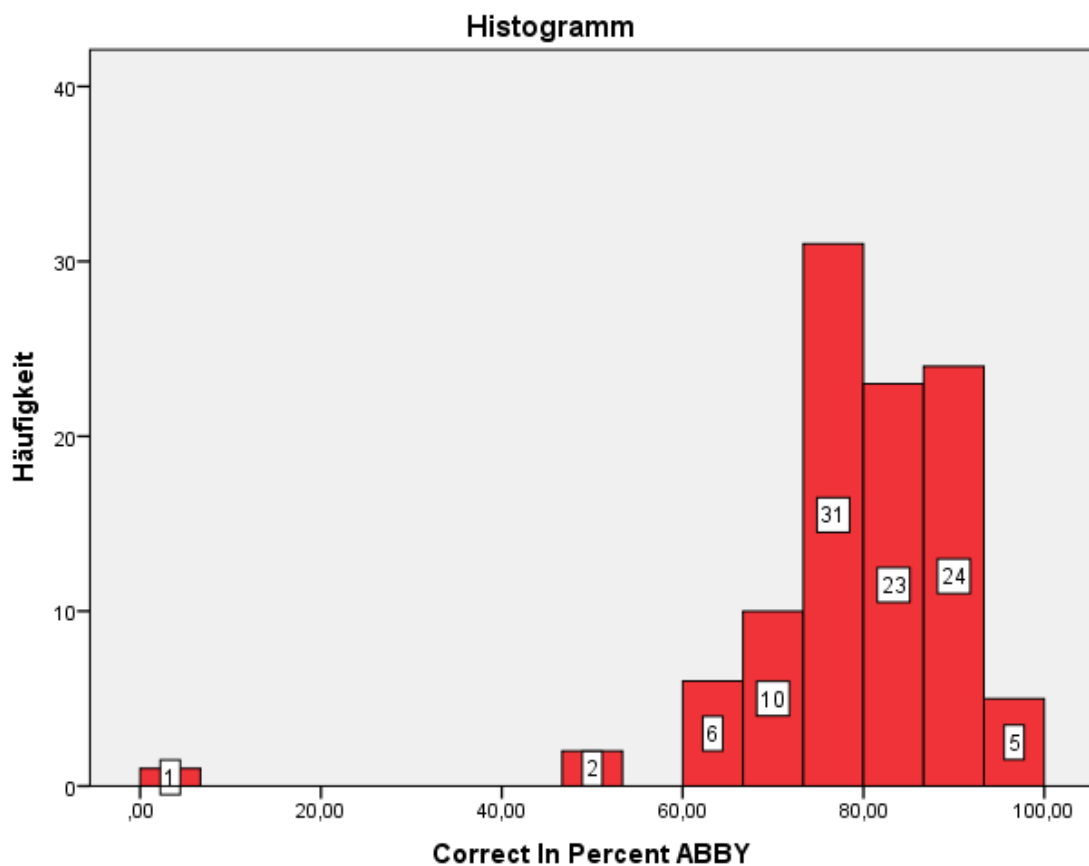
3.2 ABBY – Deskriptive Statistik

Folgende Tabelle zeigt die Ergebnisse für die relevanten Variablen zusammengefasst auf:

Tabelle 3: Deskriptive Statistik – ABBY

Deskriptive Statistiken							
	N	Bereich	Minimum	Maximum	Mittelwert	Standardabweichung	Varianz
Correct In Percent ABBY	102	95,51	,00	95,51	79,5895	12,21101	149,109
Spurious In Percent ABBY	102	34,19	,00	34,19	8,6452	6,03799	36,457
Confused In Percent ABBY	102	41,27	,00	41,27	15,5821	7,94531	63,128
Lost In Percent ABBY	102	100,00	,00	100,00	4,8285	11,19534	125,336
CER ABBY	102	91,68	8,32	100,00	29,0554	14,62504	213,892
Precision ABBY	102	91,86	,00	91,86	73,5722	12,92064	166,943
Gültige Anzahl (listenweise)	102						

Mit einem Mittelwert von 80% für die korrekte Erkennung fällt ABBY gemäß der Einstufung von Holley (2009) in die Kategorie „Poor OCR Accuracy“.

**Abbildung 9: Histogramm – ABBY Correct In Percent**

Im Histogramm erkennt man, dass die Mehrzahl der Blätter im Bereich von 60 – 90 Prozent korrekt erkannt werden. Einige wenige Blätter werden überhaupt nicht erkannt.

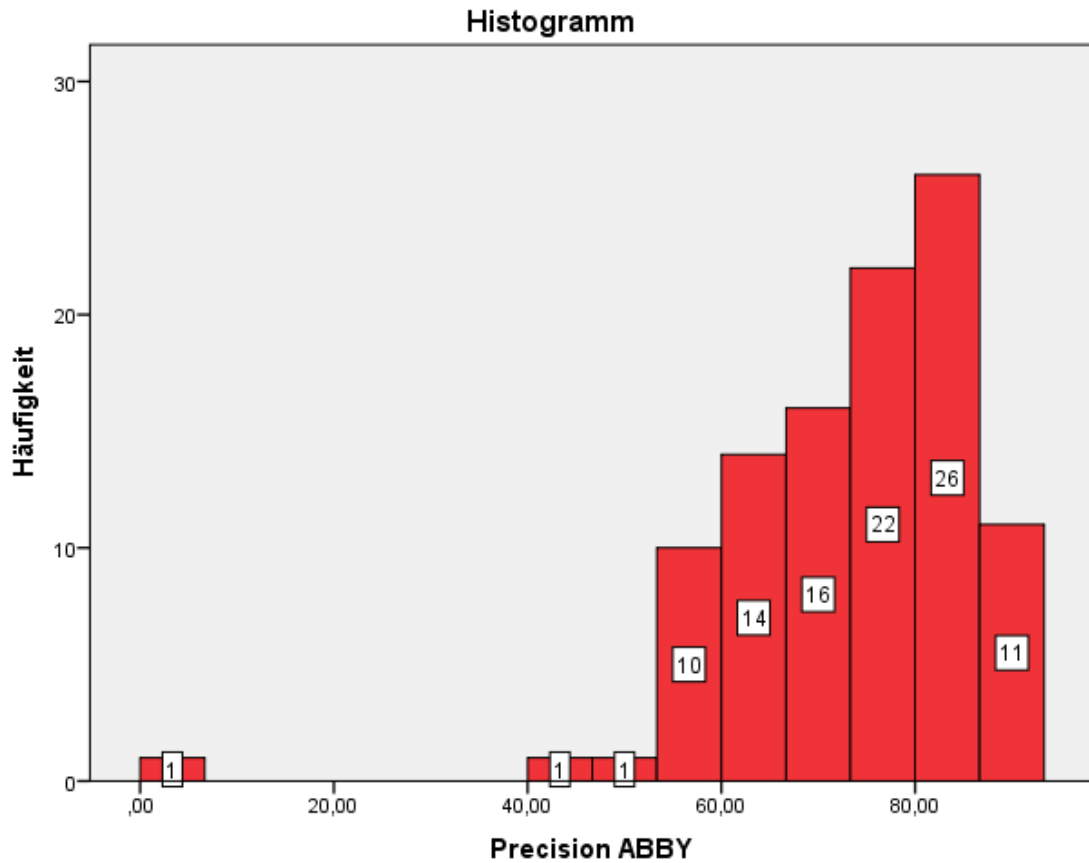


Abbildung 10: Histogramm – ABBY Precision

Die Precision hat einen Mittelwert von 74%. Dies ist eine Abweichung von der Accuracy von 6% (Correct In Percent = 80%). Dementsprechend kann man annehmen, dass die Noise nur einen geringen Einfluss auf die Performanz von ABBY hat. Lediglich die Varianz ist etwas größer.

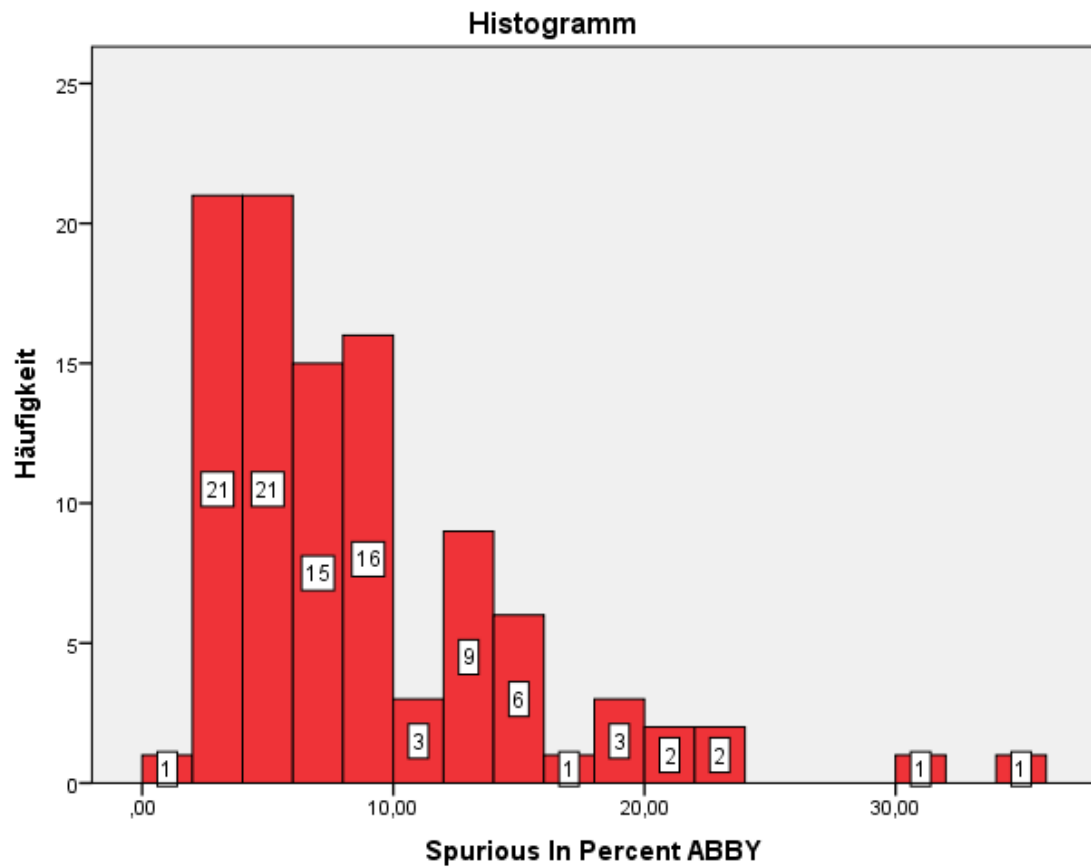


Abbildung 11: ABBY – Spurious in Percent

Die Werte für Spurious verteilen sich vor allem im Bereich zwischen 0 und 10%. Es gibt einige wenige Ausreißer im Bereich ab 20%. Der Mittelwert ist 9%.

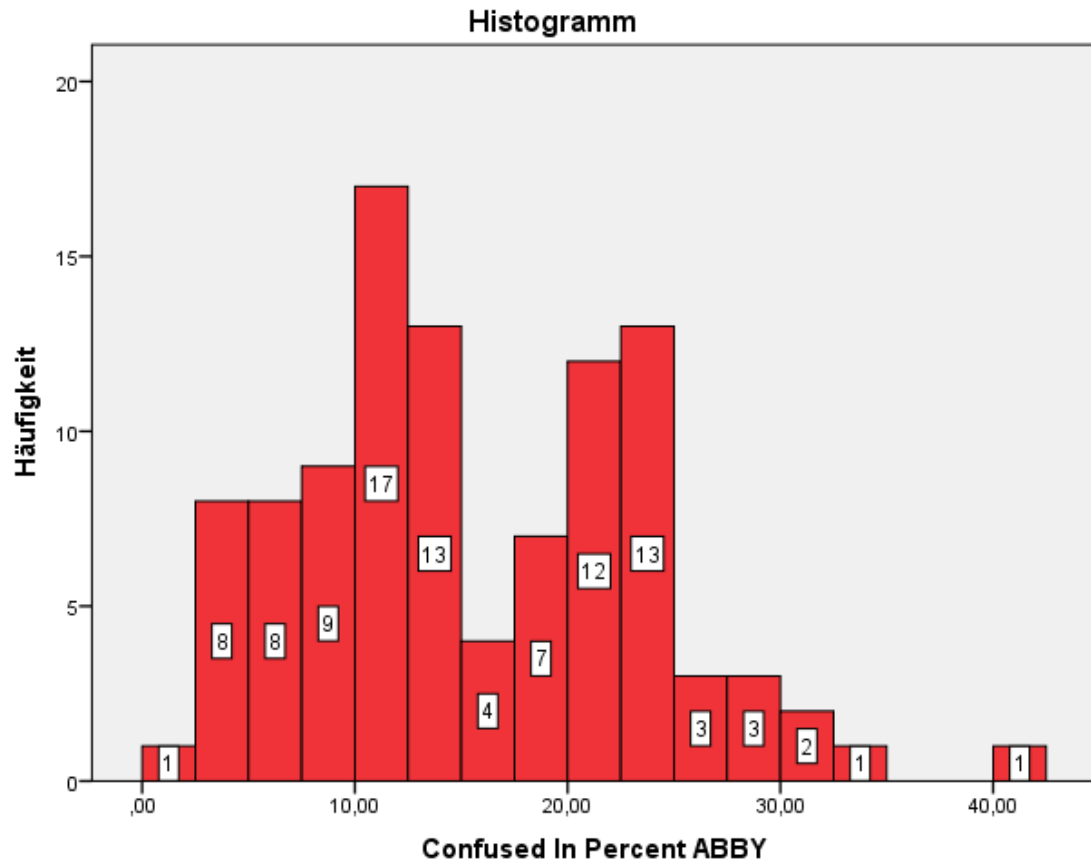


Abbildung 12: Histogramm – ABBY Confused In Percent

Die Werte für Confused-Raten streuen stark bis zum Maximum von 41%. Mit einem Mittelwert von 16% handelt es sich um die häufigste Fehlerart. Dies zeigt auf, dass ABBY (vor allem im Vergleich zu den anderen Tools) weniger Schwierigkeiten mit der allgemeinen Erkennung der Existenz eines Zeichens hat. Also die Fehlertypen Lost und Spurious weniger häufig auftreten.

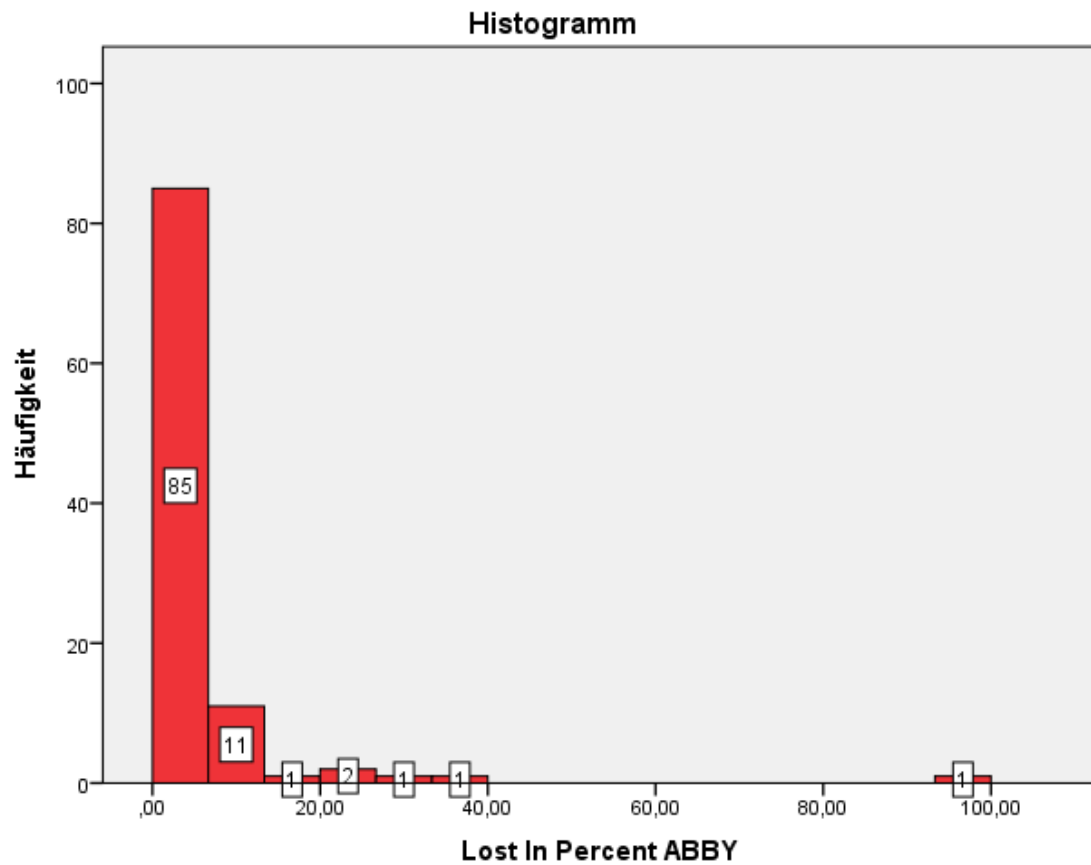


Abbildung 13: Histogramm – ABBY Lost In Percent

Für die große Mehrzahl der Blätter gilt, dass keine oder sehr wenige Zeichen verloren gehen. Für ein Blatt wurde der Text gar nicht erkannt. Der Mittelwert ist knapp 5%.

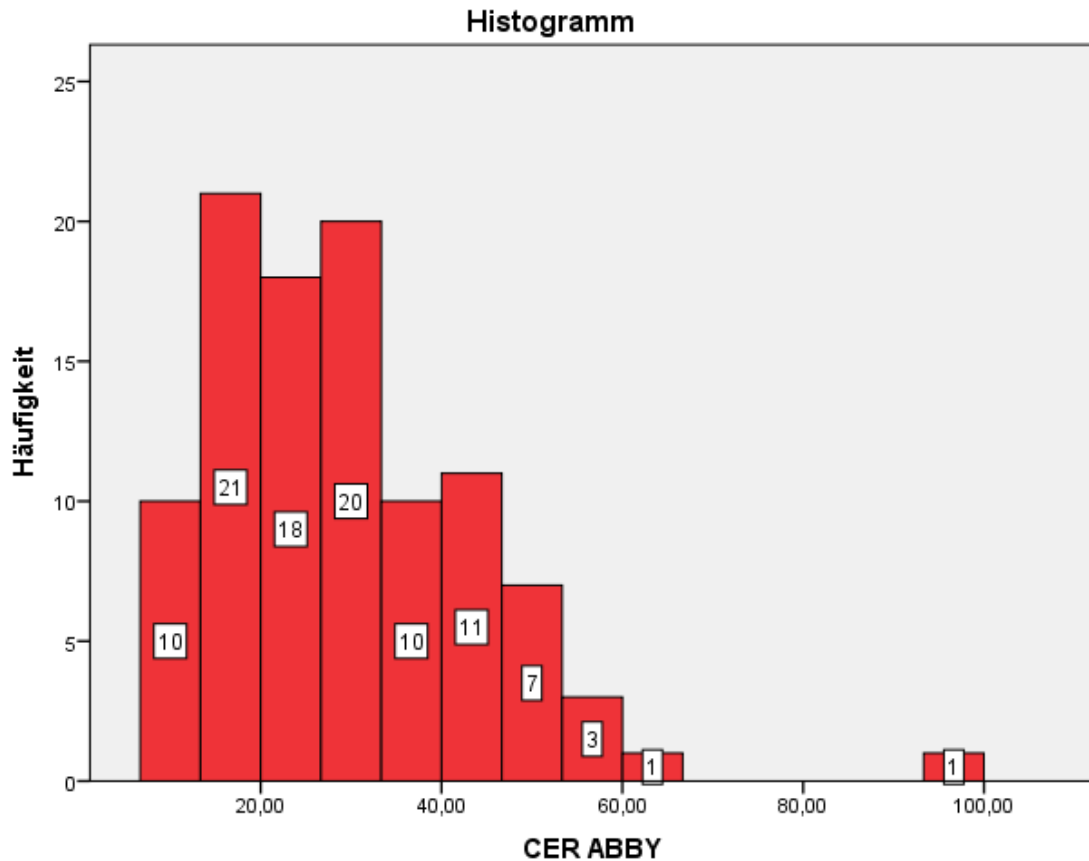


Abbildung 14: Histogramm – ABBY CER

Die Character Error Rate zeigt zwingenderweise ein ähnliches Bild wie die Genauigkeit (Correct in Percent) nur spiegelverkehrt. Mit einem Mittelwert von 29% stellt man jedoch auch hier gemäß der Einschätzung von Holley (2009) eine eher schwache OCR-Leistung fest.

Bei der Betrachtung der Verteilung von Zeichentypen gibt es drei Möglichkeiten der Analyse. Zum einen kann man alle Zeichentypen (Correct, Spurious, Confused, Lost) gesamt auffassen. Dies führt dazu, dass die Zeichenanzahl für ein Liedblatt für alle Tools unterschiedlich sein kann, da die Zahl der Spurious-Zeichen beliebig hoch sein kann. Dennoch kann man so singulär die Leistung eines Tools beurteilen. Als zweites kann man die Fehlertypen nur in Bezug auf den Grounded-Truth-Text betrachten. Dadurch fallen die Spurious-Werte weg. Dies macht Vergleichbarkeit möglich, da die totale Zeichenanzahl für jeden Grounded-Truth-Text gleich ist. Als letztes kann man die Verteilung in Bezug auf den reinen OCR-Output betrachten. Dadurch werden die Lost-Zeichen ignoriert. Die unterschiedliche Aussagekraft aller Grafiken ist im Folgenden zu beachten. Die erst genannte Grafik hat natürlich die deutlichste Aussagekraft und sollte

primär beachtet werden. Die weiteren sind lediglich spezielle Variationen in Bezug zu einem konkreten Text (Grounded Truth und OCR-Output).

Folgendes Kreisdiagramm zeigt die Verteilung der absoluten Werte aller möglichen Ausprägungen eines Zeichens (Correct, Spurious, Confused, Lost) über den gesamten Korpus hinweg. Oben sieht man dabei den prozentualen Anteil und darunter die absolute Zahl. Beispielsweise gibt ABBY in seinem gesamten Ausgabe-Korpus 74% korrekter Zeichen aus (exakt 41 667).

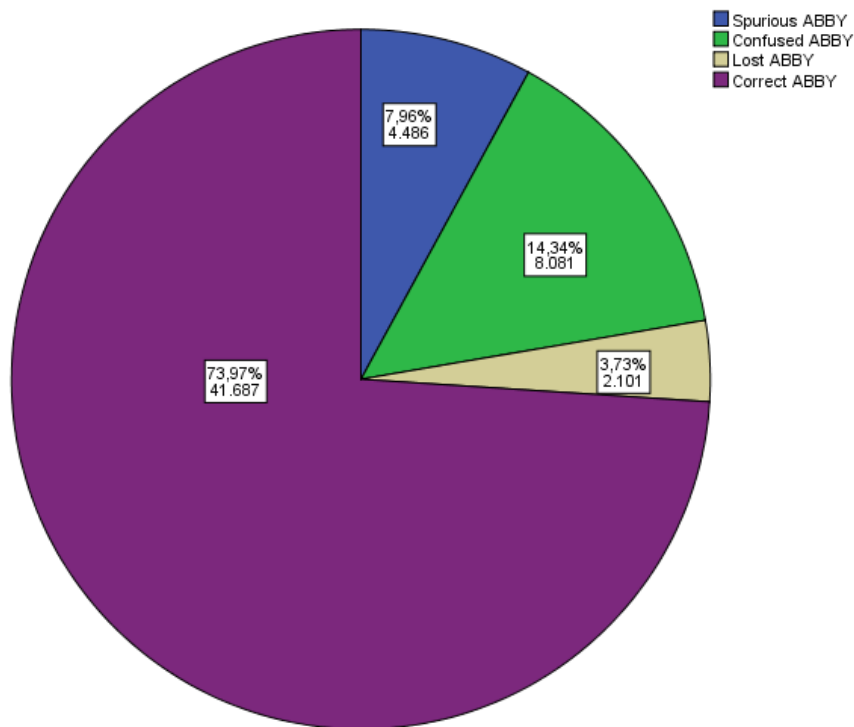


Abbildung 15: Kreisdiagramm – ABBY Complete

Bezogen auf die Grounded-Truth-Daten, also ohne Betrachtung von Spurious-Zeichen, sieht das Kreisdiagramm folgendermaßen aus:

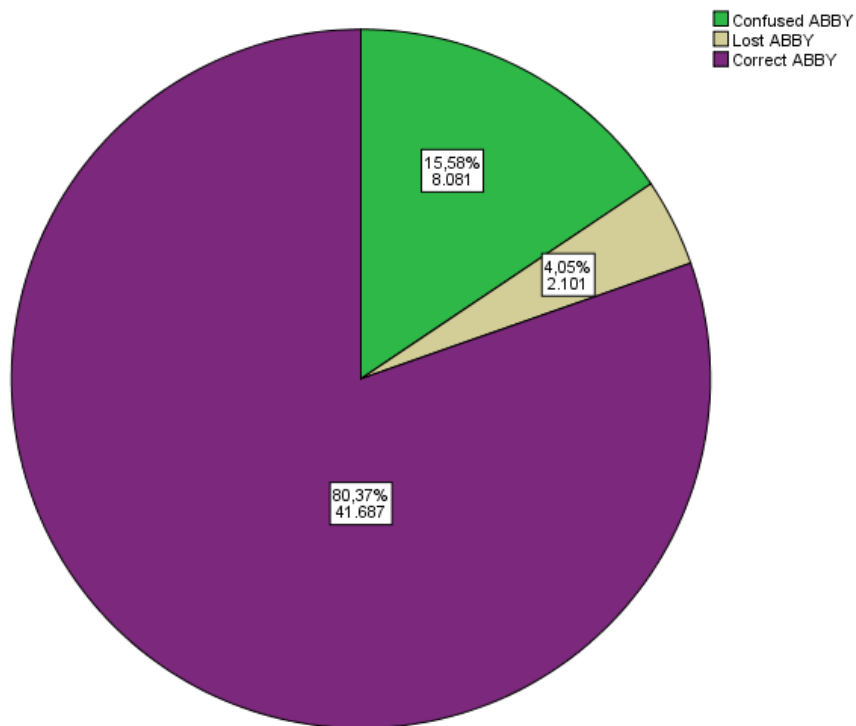


Abbildung 16: Kreisdiagramm – ABBY Grounded Truth

Bezogen auf den reinen OCR-Output müssen die Lost-Zeichen weggelassen werden.

Dieses Kreisdiagramm zeigt die Verteilung im reinen OCR-Text.

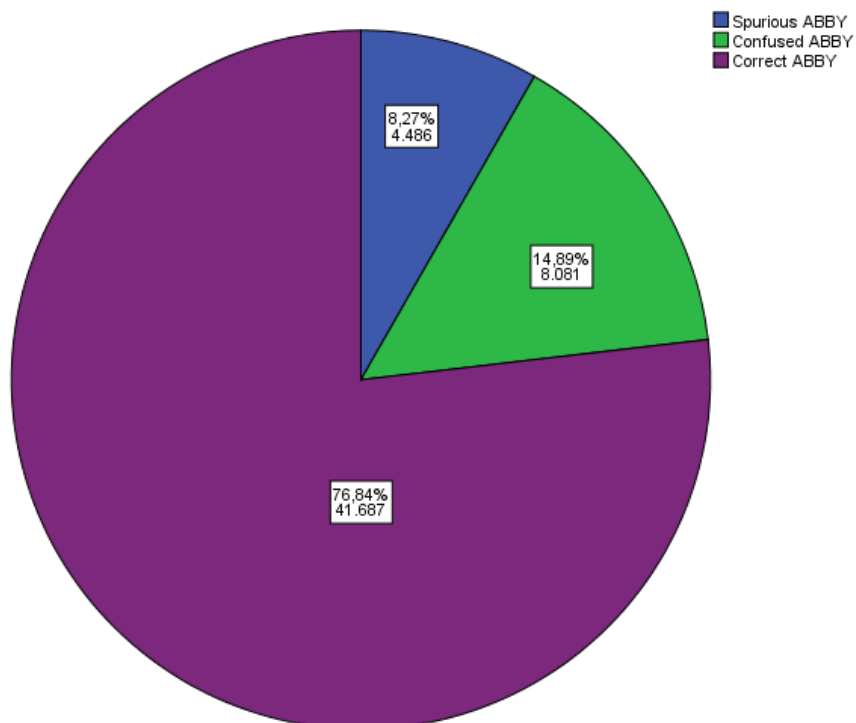


Abbildung 17: Kreisdiagramm – ABBY OCR-Output

3.3 Omnipage Professional – Deskriptive Statistik

Diese SPSS-Tabelle fasst alle Ergebnisse der deskriptiven Statistik zusammen:

Tabelle 4: Deskriptive Statistik – Omnipage

Deskriptive Statistiken							
	N	Bereich	Minimum	Maximum	Mittelwert	Standardabweichung	Varianz
Correct In Percent Omnipage	102	90,11	,00	90,11	55,7934	27,30632	745,635
Spurious In Percent Omnipage	102	182,46	,00	182,46	8,0192	19,46479	378,878
Confused In Percent Omnipage	102	78,66	,00	78,66	19,9421	14,81493	219,482
Lost In Percent Omnipage	102	99,37	,63	100,00	24,2645	29,52750	871,873
CER Omnipage	102	240,23	10,65	250,88	52,2256	33,09877	1095,528
Precision Omnipage	102	89,43	,00	89,43	52,0811	25,91975	671,834
Gültige Anzahl (listenweise)	102						

Mit einer durchschnittlichen Erkennungsrate von 56% ist auch Omnipage gemäß der Klassifikation von Holley (2009) weit unter tolerabler Qualität.

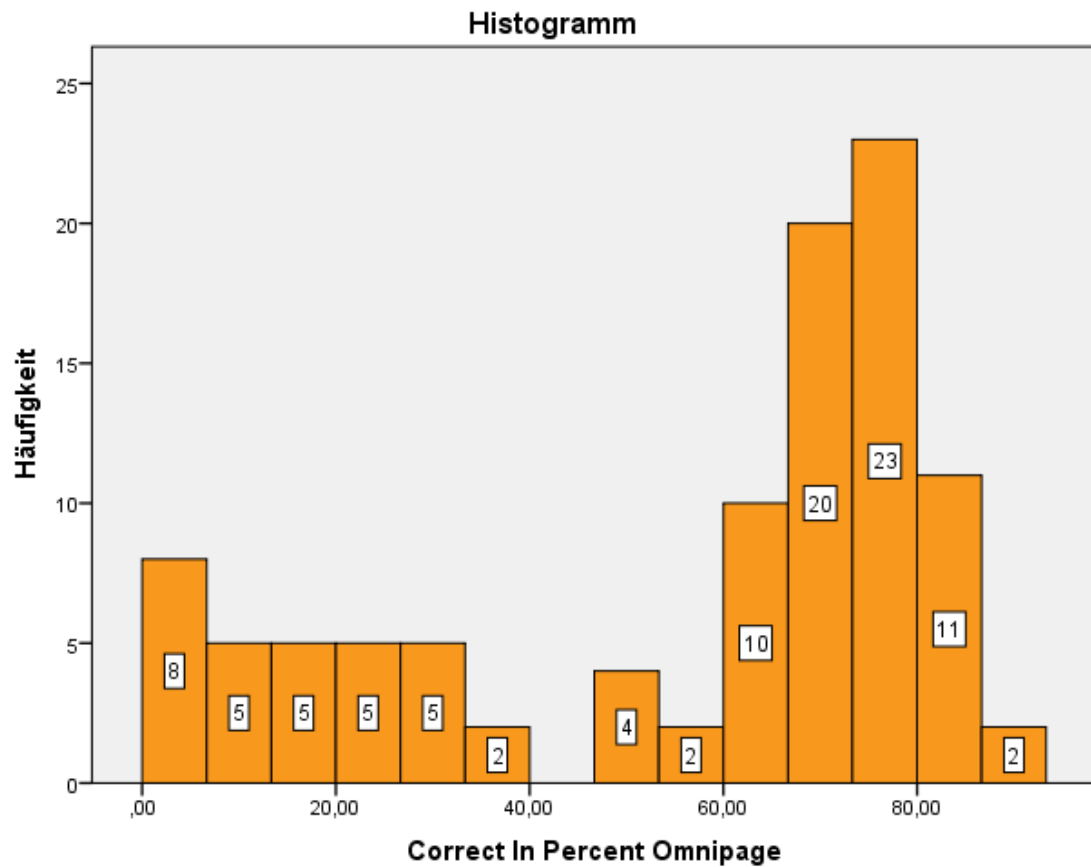


Abbildung 18: Histogramm – Omnipage Correct In Percent

Die Erkennungsrate von Omnipage streut stark für die Liedblätter mit einem geringfügigen Schwerpunkt zwischen 60 und 80%. 30 Blätter werden unter 40% erkannt, 8 Liedblätter überhaupt nicht.

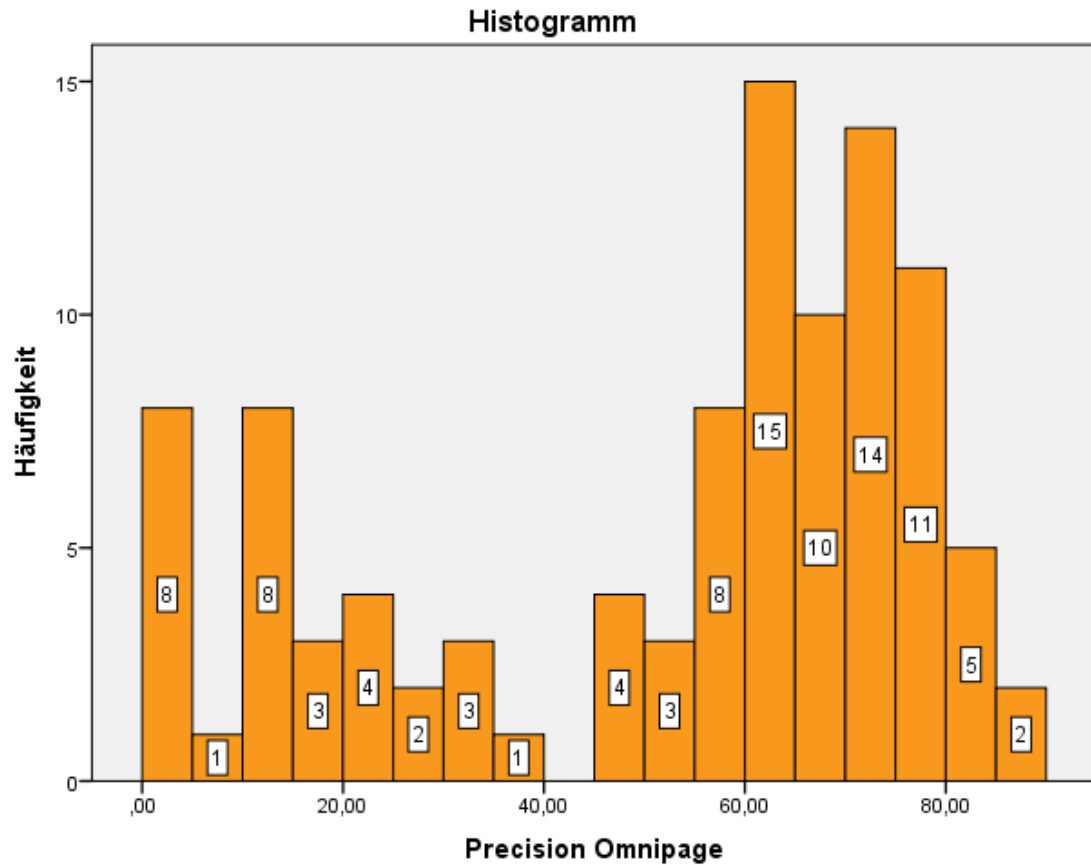


Abbildung 19: Histogramm – Omnipage Precision

Für die Precision-Werte liegt auch ein ähnliches Bild vor, nur geringfügig schlechter. Auch hier haben ca. 30 Blätter eine Precision unter 40 %. Der Mittelwert liegt bei 52% und auch die Standardabweichung ist ähnlich groß. Die Noise ist folglich kein großer Einflussfaktor auf die Performanz des Tools.

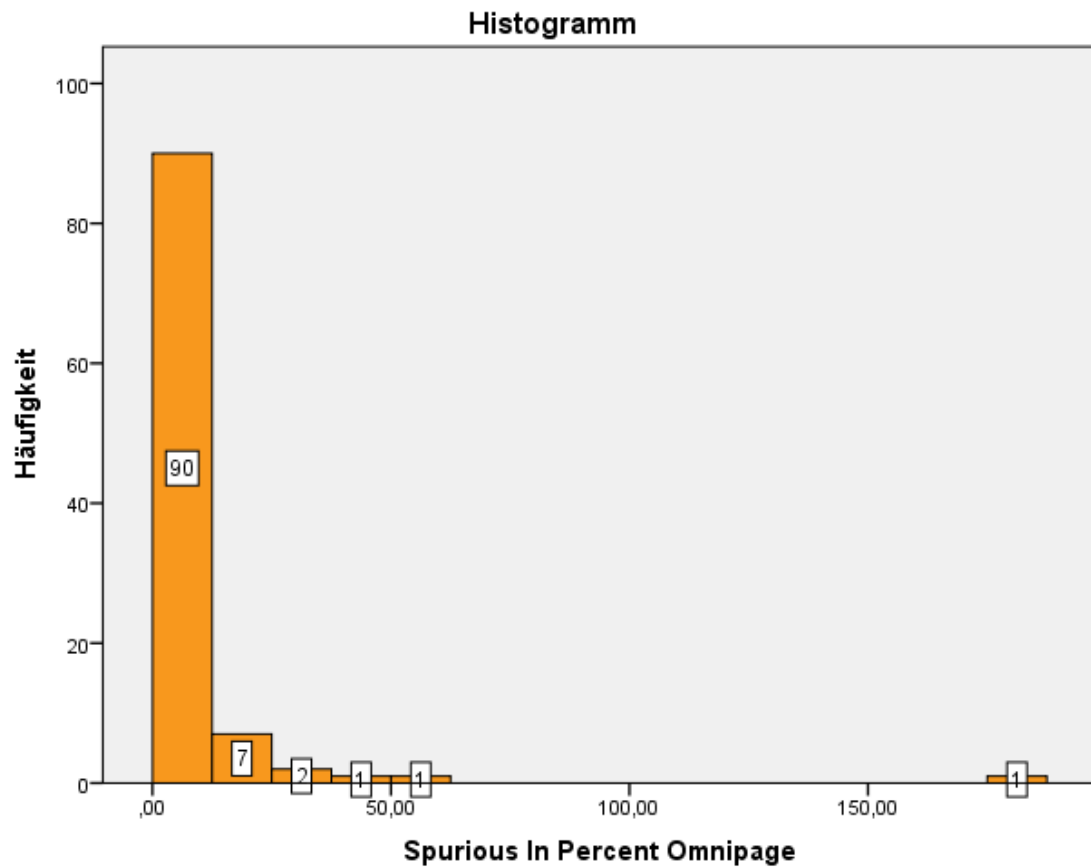


Abbildung 20: Histogramm – Omnipage Spurious In Percent

Die große Mehrzahl der Blätter erzeugt keine oder wenig Noise bei Omnipage. Eine mittelgroße Anzahl von Blättern erzeugt einen Spurious-Anteil von 20 – 50%. Ein Ausreißer liefert einen Spurious-Wert von 183%. Mit einem Mittelwert von 8% ist Omnipage in Bezug auf diese Fehlerart ähnlich zu ABBY.

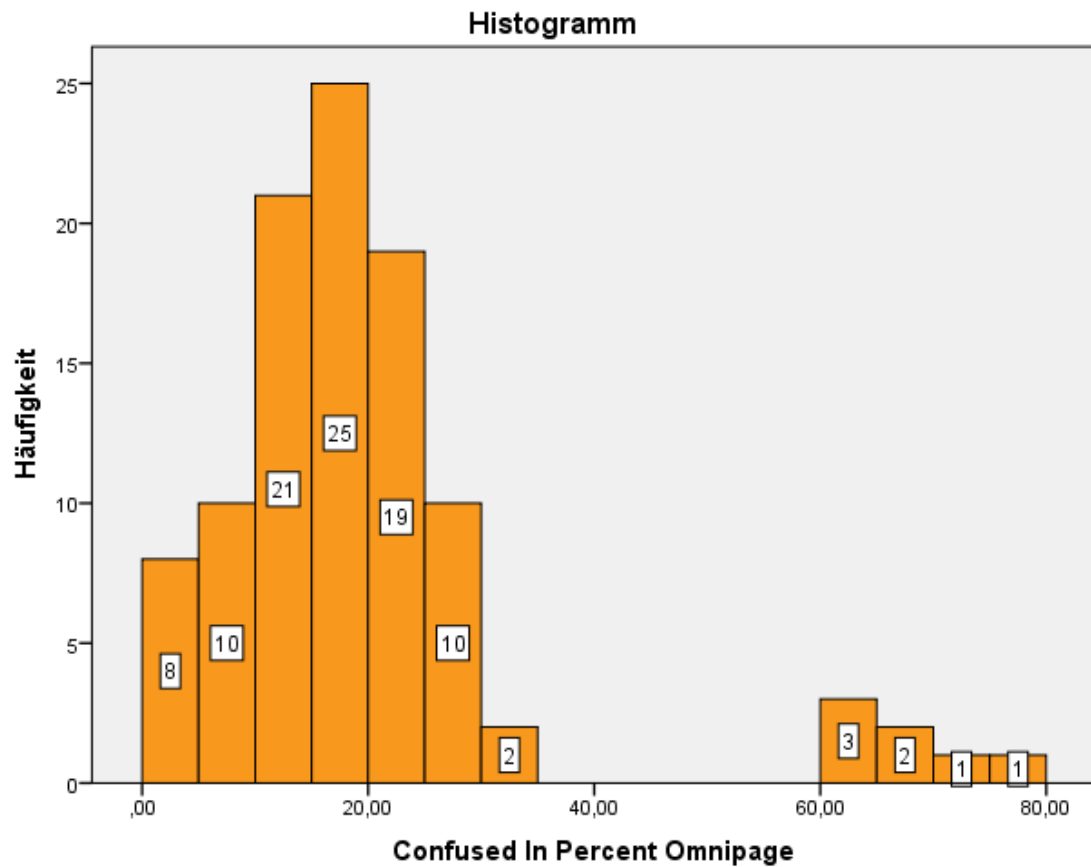


Abbildung 21: Histogramm – Omnipage Confused In Percent

Der Mittelwert der Confused-Raten liegt bei ca. 20%. Die Mehrzahl der Blätter haben Raten zwischen 0 und 30%. Sieben Blätter haben Raten über oder gleich 60%.

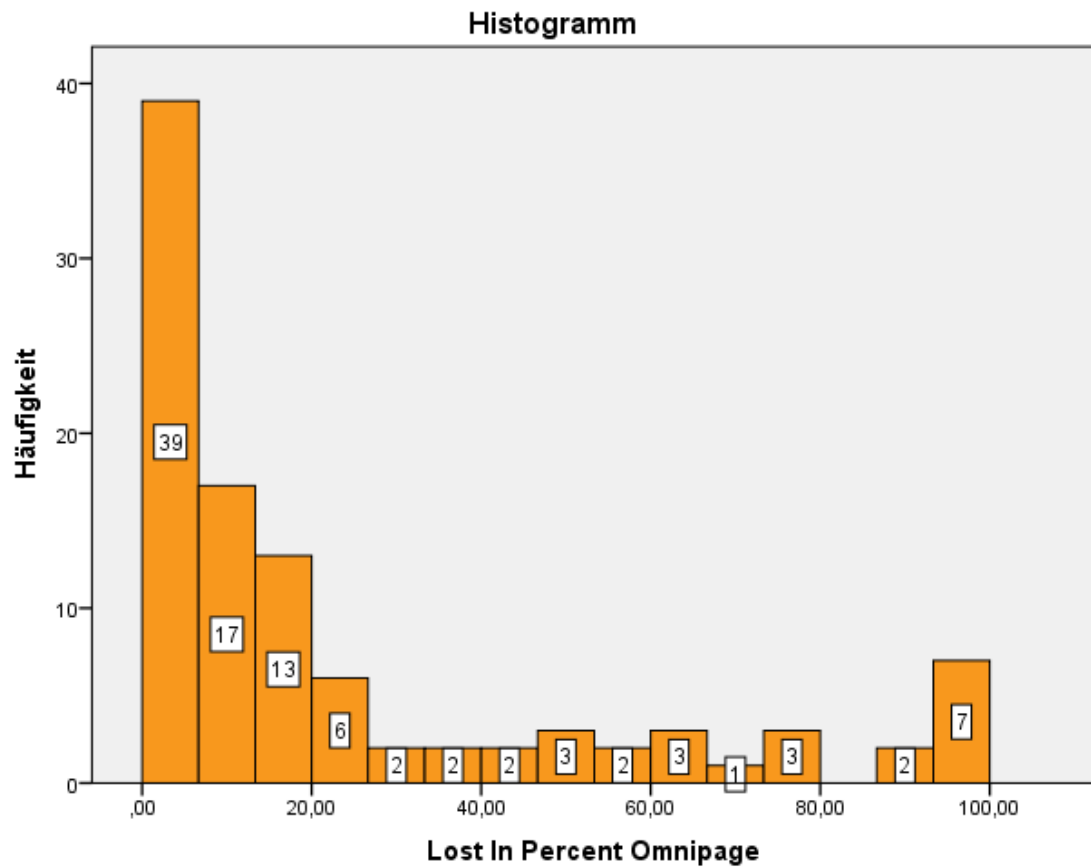


Abbildung 22: Histogramm – Omnipage Lost In Percent

Der Hauptanteil der Lost-Werte für Liedblätter verläuft im Bereich von 0 – 20% konstant abnehmend. Ab 20% verteilen sich die Werte stark. Sieben Blätter werden überhaupt bzw. fast gar nicht erkannt. Der Mittelwert beträgt 24%.

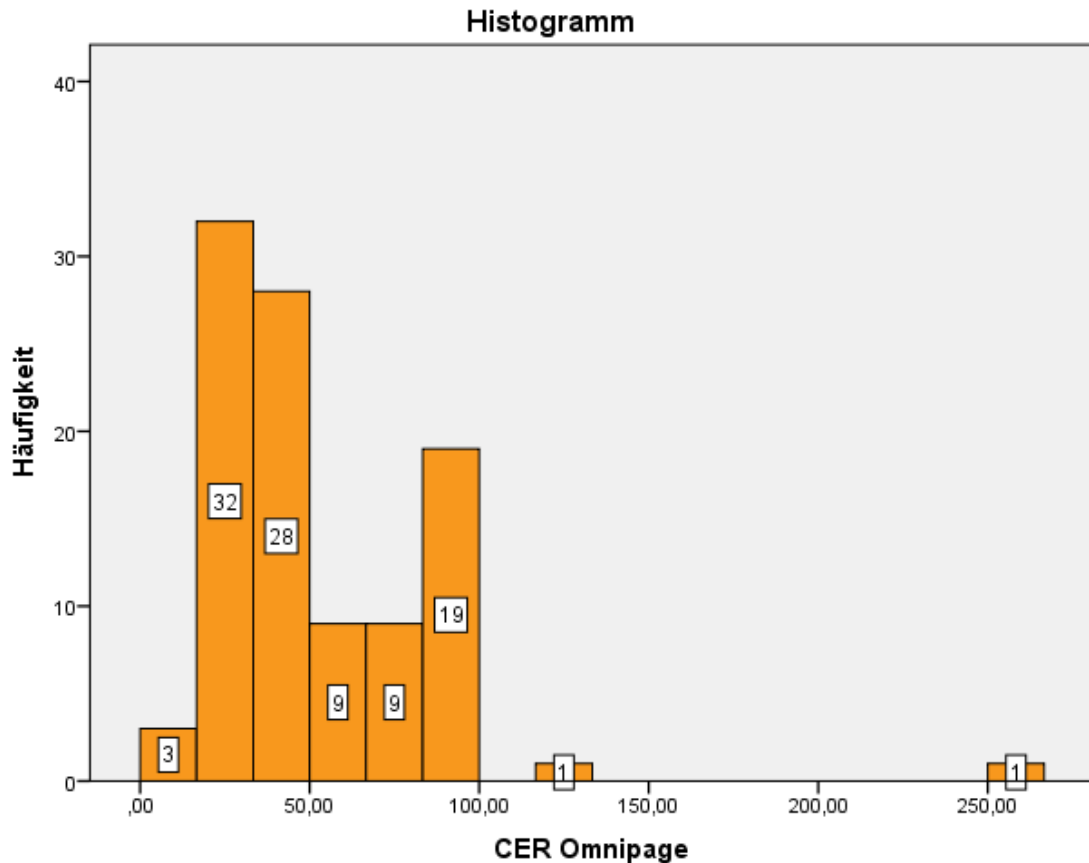


Abbildung 23: Histogramm – Omnipage CER

Die CER-Rate streut zwischen 0 und 100%. Aufgrund vereinzelter großer Spurious-Raten erhält man bei Omnipage einige Raten über 100%. Mit einem Mittelwert von 53% ist die Fehlerproduktion sehr hoch. Der Wert sagt aus, dass im Schnitt die Hälfte des OCR-Outputs von Omnipage ausgebessert oder ergänzt werden muss.

Analog zur deskriptiven Statistik von ABBY sollen die folgenden Kreisdiagramme die Häufigkeit von Zeichentypen (Correct, Spurious, Confused, Lost) gesamt, für den OCR-Output als auch in Bezug auf den Grounded-Truth-Text, über den ganzen Korpus aufzeigen:

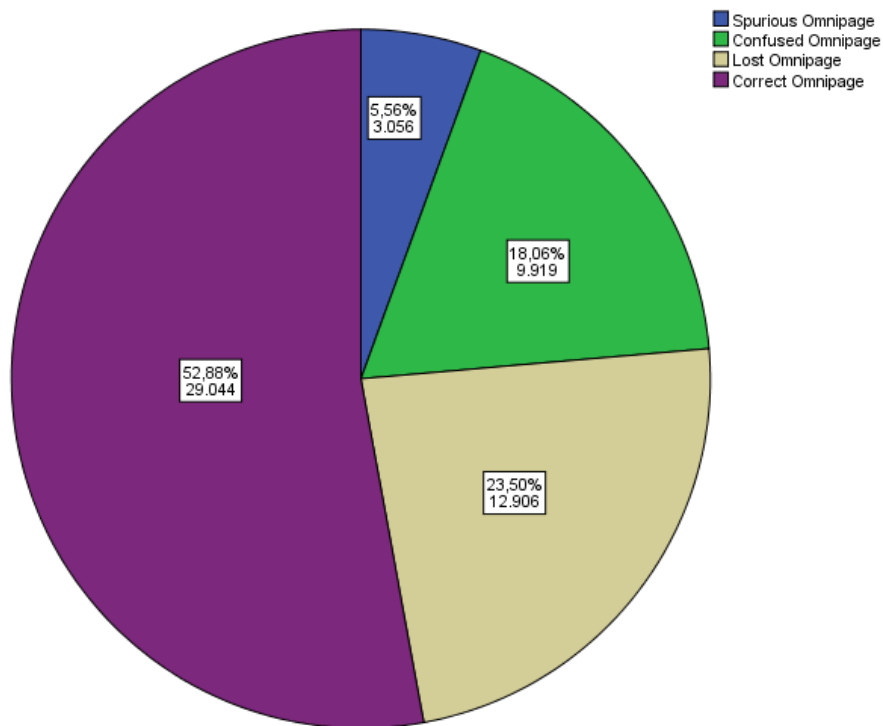


Abbildung 24: Kreisdiagramm – Omnipage Complete

Man erkennt, dass die häufigste Fehlerart, die ist, dass Zeichen gar nicht erkannt werden (23,5%). Die Hälfte des Outputs ist entweder falsch (Spurious, Confused) oder muss ergänzt werden. Mit ca. 6% ist der Anteil von Noise verhältnismäßig gering.

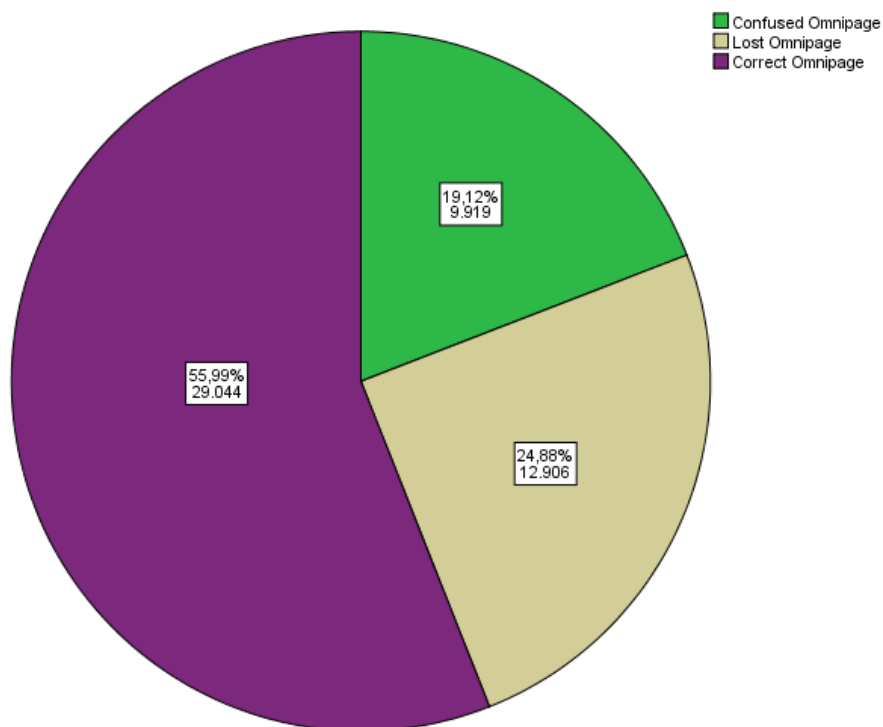


Abbildung 25: Kreisdiagramm – Omnipage Grounded Truth

In Bezug auf den Grounded-Truth-Text ist das Ergebnis aufgrund der geringen Spurious-Werte ähnlich zu oben. Fast ein Viertel der Zeichen wird nicht erkannt. Annähernd ein weiteres Viertel wird falsch erkannt.

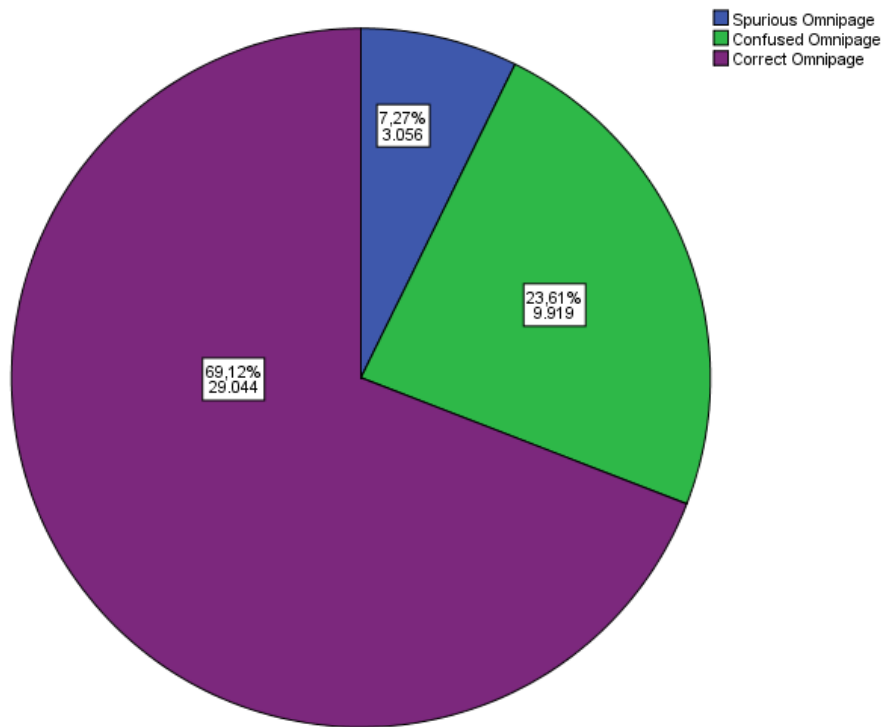


Abbildung 26: Kreisdiagramm – Omnipage OCR-Output

Da bezogen auf den OCR-Output Lost-Zeichen nicht angegeben sind, erscheint das Ergebnis in diesem Kreisdiagramm positiver als die reale Leistung ist.

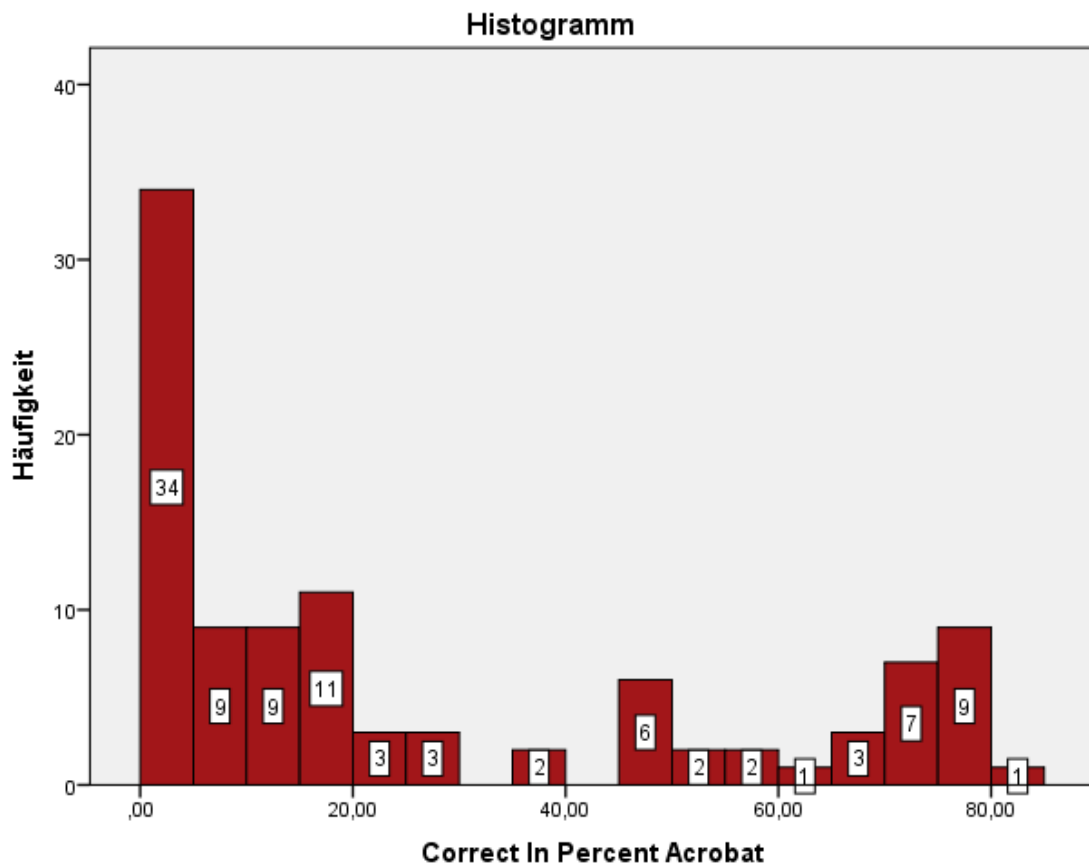
3.4 Adobe Acrobat X Pro – Deskriptive Statistik

Eine Zusammenfassung der deskriptiven Daten in Bezug auf die wichtigsten Performanzparameter:

Tabelle 5: Deskriptive Statistik – Acrobat X Pro

Deskriptive Statistiken							
	N	Bereich	Minimum	Maximum	Mittelwert	Standardabweichung	Varianz
Correct In Percent Acrobat	102	82,61	,00	82,61	26,3960	28,29132	800,399
Spurious In Percent Acrobat	102	407,89	,00	407,89	19,0471	59,47921	3537,777
Confused In Percent Acrobat	102	72,00	,00	72,00	15,7189	16,59953	275,544
Lost In Percent Acrobat	102	100,00	,00	100,00	57,8852	40,02903	1602,323
CER Acrobat	102	434,65	25,00	459,65	92,6514	58,58258	3431,919
Precision Acrobat	102	76,77	,00	76,77	21,9626	24,02734	577,313
Gültige Anzahl (listenweise)	102						

Die Erkennungsrate (Correct In Percent) ist nach Holley (2009) weit unter dem Zielwert (90%). Es lässt sich leicht erkennen, dass das Tool im Vergleich zu den anderen Tools und in Bezug auf jede Variable deutlich schlechtere Ergebnisse ausgibt.

**Abbildung 27: Histogramm – Acrobat Correct In Percent**

Der Großteil der Blätter ergibt Erkennungsraten zwischen 0 und 20%. 21 Blätter, also etwa ein Fünftel des Korpus hat eine Erkennungsrate von über 60%. Die Werte streuen wenig, der Mehranteil der Blätter hat sehr geringe Correct-Werte.

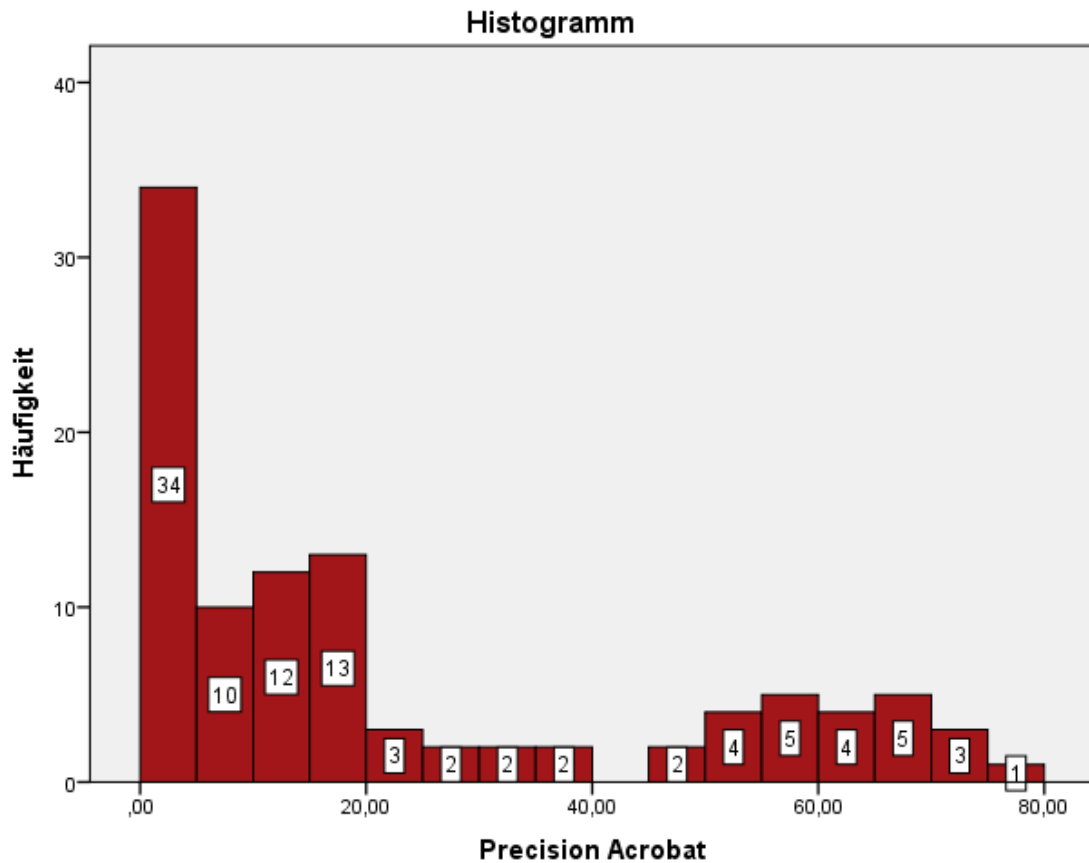


Abbildung 28: Histogramm – Acrobat Precision

Auch bei diesem Tool zeigt die Auswertung der Precision-Daten, dass kein wesentlicher Einfluss durch die überflüssig hinzugefügten Zeichen zustande kommt. Die Precision-Werte sind geringfügig schlechter als die Accuracy-Werte. Insgesamt weist aber auch diese Metrik auf eine schlechte Performanz hin. 57 Bilder haben eine Precision < 20%, d.h. mehr als 80% der Zeichen dieser OCR-Outputs muss in irgendeiner Form ausgebesert werden.

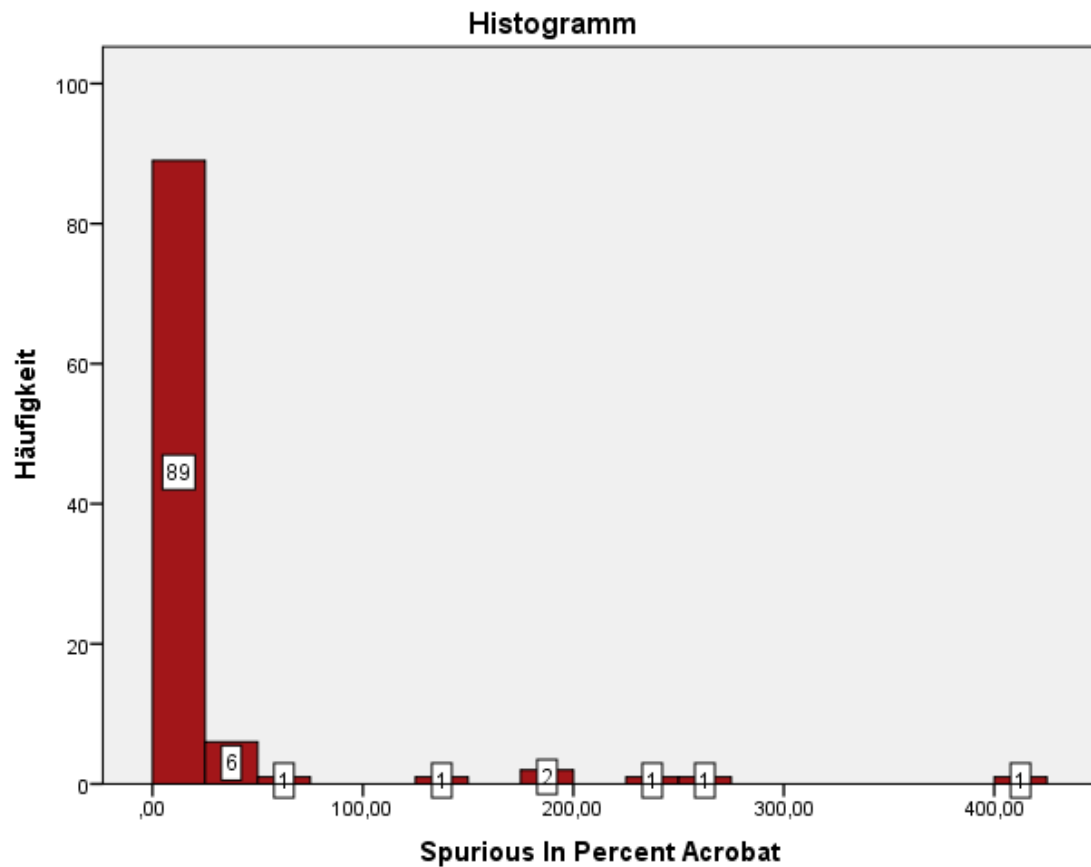


Abbildung 29: Histogramm – Acrobat Spurious In Percent

Die Spurious-Werte sind im Vergleich zu anderen Fehlertypen mit einem Mittelwert von 19% im Vergleich zu den anderen Fehlertypen recht gering, im Verhältnis zu den anderen Tools sehr groß. Sechs Dokumente haben extrem großen Noise-Anteil mit über 100%.

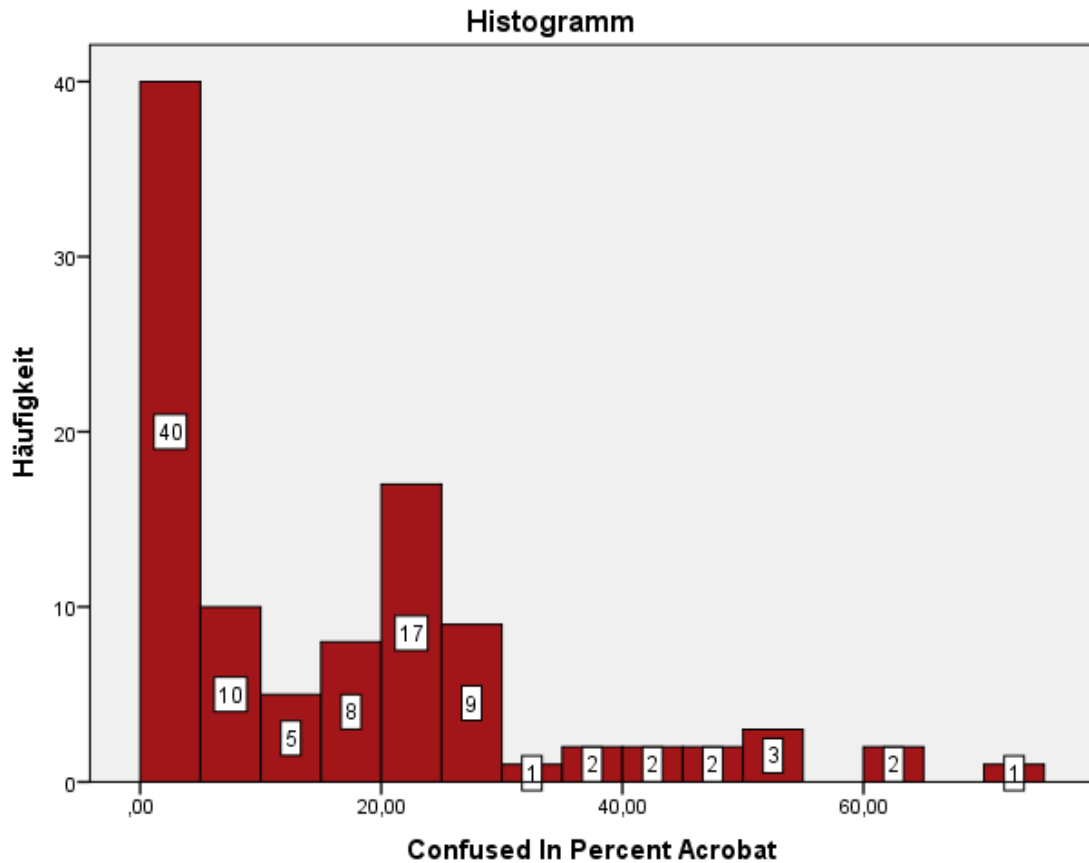


Abbildung 30: Histogramm – Acrobat Confused In Percent

Bezüglich der Confused-Werte lässt sich feststellen, dass der Großteil der Blätter, nämlich 40, Werte zwischen 0 und 5 Prozent besitzt. Generell ist der Kernbereich im Intervall zwischen 0 und 30% zu sehen. Danach hat man vereinzelte Blätter bis hin zum Maximum von 72%. Das arithmetrische Mittel beträgt ca. 16%. In Bezug auf die Confused-Werte ist demnach das Tool ähnlich „gut“ wie ABBY.

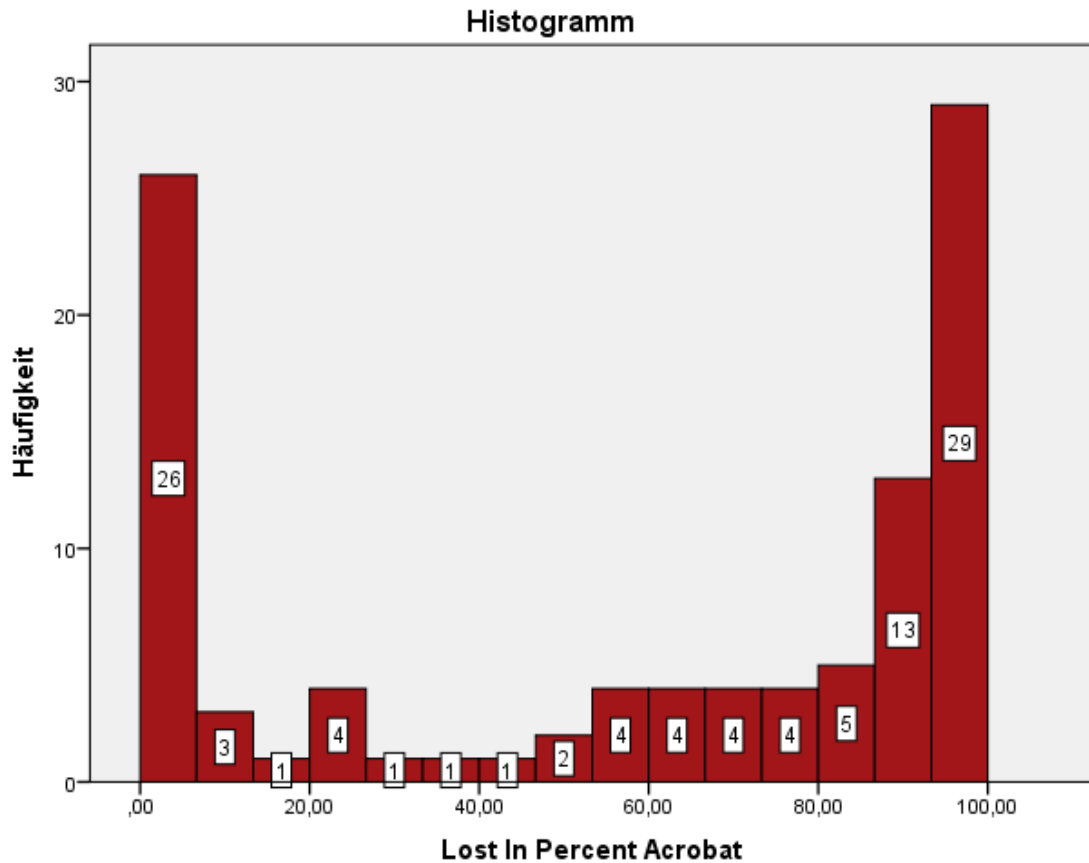


Abbildung 31: Histogramm – Acrobat Lost In Percent

Das große Problem von Acrobat X Pro erkennt man an den Lost-Werten. Tatsächlich folgt die Verteilung einer U-Verteilung. Mit knapp 30 Blättern über der Rate von 90% bedeutet das, dass fast ein Drittel des Korpus nicht erkannt wurde. Der Mittelwert mit 58% besagt, dass im Schnitt mehr als die Hälfte der Zeichen des Korpus nicht erkannt wurden.

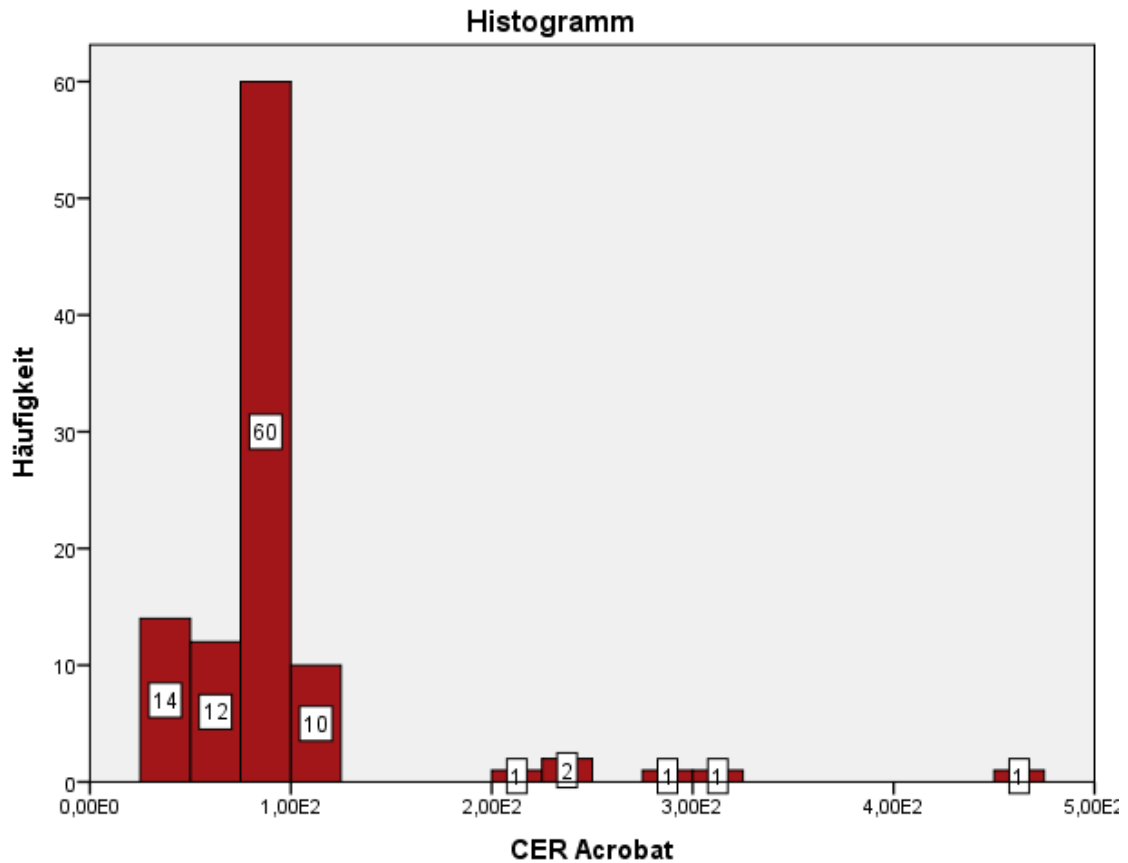


Abbildung 32: Histogramm – Acrobat CER

Auch bei den CER-Werten sorgen einige Extremwerte über 100% für eine Verzerrung. Mit einem Minimum von 25% sieht man jedoch, dass das am besten erkannte Blatt trotzdem einen Fehleroutput von genau einem Viertel des Grounded-Truth-Text erzeugt (bezogen auf alle Fehlertypen). 76 Blätter haben einen CER-Wert über 75%.

Analog zu den obigen Tools seien hier noch Kreisdiagramme angegeben um die Fehlerverteilung zu verdeutlichen:

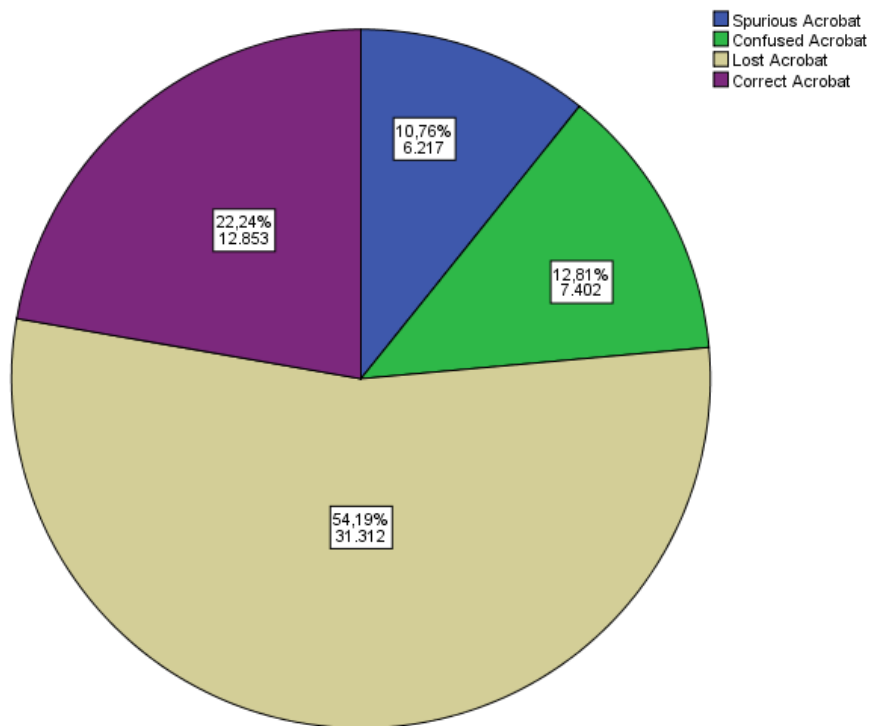


Abbildung 33: Kreisdiagramm – Acrobat Complete

Der Großteil der Zeichen geht einfach verloren. Nur ein Viertel wird korrekt erkannt.

Bezogen auf den reinen Grounded-Truth-Korpus ist das Verhältnis ähnlich:

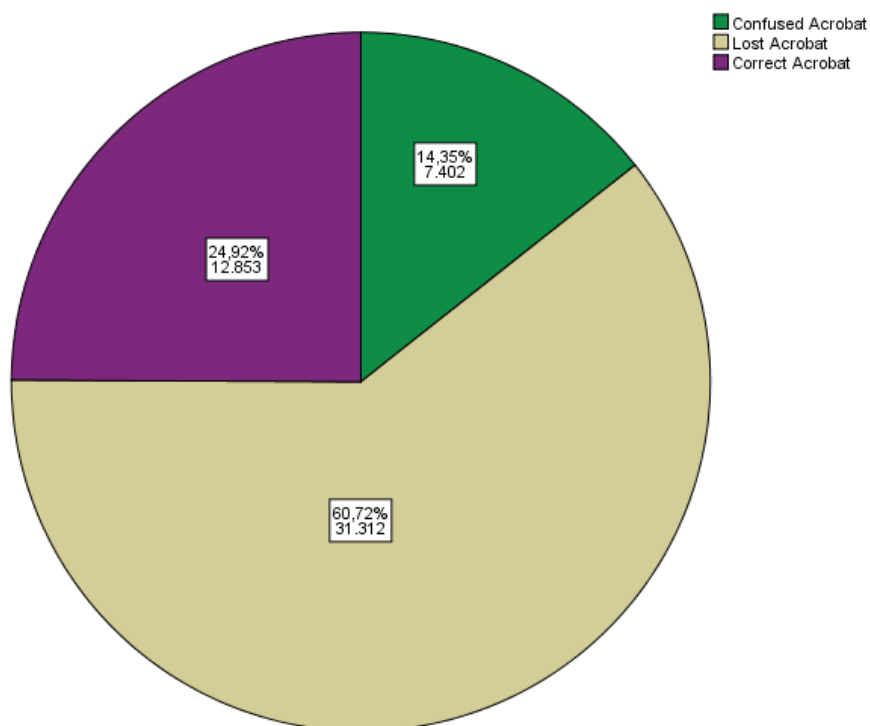


Abbildung 34: Kreisdiagramm – Acrobat Grounded Truth

Bezogen auf den OCR-Text kaschiert die fehlende Angabe von Lost-Zeichen natürlich das negative Endergebnis.

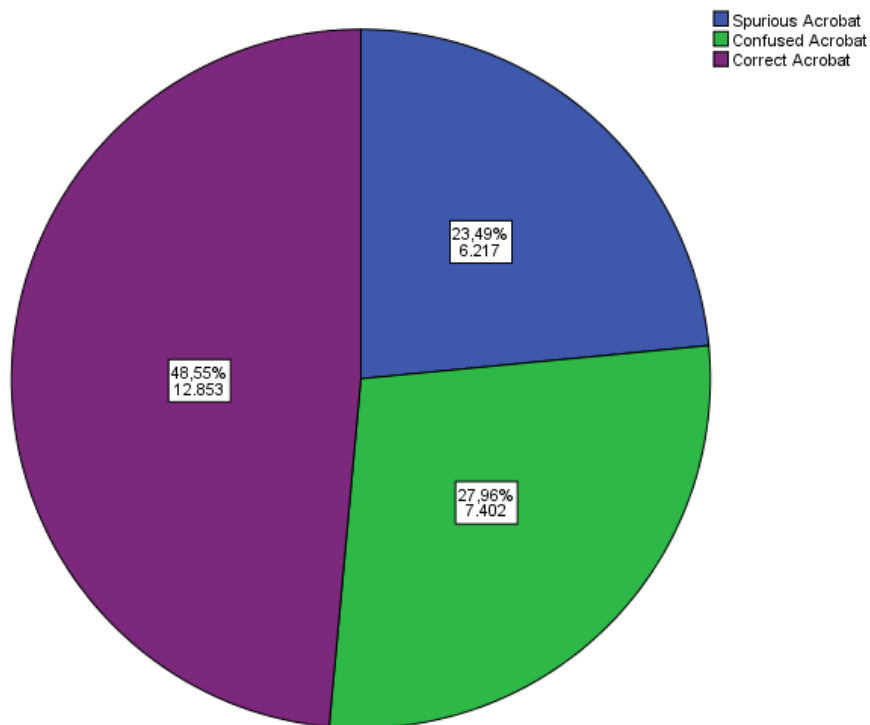


Abbildung 35: Kreisdiagramm – Acrobat OCR-Output

Aber selbst nur im OCR-Output ist mehr als die Hälfte der dort vorhandenen Zeichen falsch oder überflüssig.

3.5 Vergleich – Deskriptive Statistik und Inferenzstatistik

3.5.1 Hypothesenformulierung

Nach Interpretation der bisherigen deskriptiven Statistik kann man folgende Alternativhypothese aufstellen mit den Tools sowie dem Begriff der Leistung, operationalisiert über die Variablen Correct In Percent (je größer desto besser), Precision, Spurious In Percent, Confused In Percent, Lost In Percent, CER (je kleiner desto besser). Der Begriff des Testkorpus wird in 2.1 genauer beschrieben:

H1: *Es gibt einen Unterschied in der Performanz der OCR-Tools ABBY, Omnipage und Adobe Acrobat X Pro für den Testkorpus.*

Diese Hypothese zerfällt bei genauerer Analyse des Datensatzes in unterschiedliche entgegengesetzt gerichtete Einzelhypothesen:

H1.1: *ABBY hat eine bessere Performanz als Omnipage für den Testkorpus.*

H1.2: *ABBY hat eine bessere Performanz als Adobe Acrobat X Pro für den Testkorpus.*

H1.3: *Omnipage hat eine bessere Performanz als Adobe Acrobat X Pro 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 OCR-Tools ABBY, Omnipage und Adobe Acrobat X Pro für den Testkorpus.*

Es ist zu beachten, dass die obigen Hypothesen für jede Performanz-Variable noch mal in weitere Einzelhypothesen zerfallen. Auch ist die Angabe der Richtung dieser Variable, um „besser“ zu sein, notwendig. Die konkreten Hypothesen verlaufen demnach nach folgendem Verfahren:

H1.1 (Correct In Percent): *ABBY hat eine bessere Erkennungsrate als Omnipage für den Testkorpus.*

H1.1 (CER): *Abby hat eine geringere CER als Omnipage für den Testkorpus.*

usw.

Von einer vollen Auflistung wird hier jedoch abgesehen. Im Folgenden wird sich immer auf die H1, H1.1, H1.2 und H1.3 im Kontext der betrachteten Variable bezogen.

3.5.2 Signifikanztest – Statistisches Vorgehen

Bei den Stichproben der jeweiligen Performanz-Variablen für jedes Tool handelt es sich um abhängige Stichproben (Leonhart, 2013, S. 190), auch verbundene oder gepaarte Stichproben genannt. Elemente von zwei Stichproben (bzw. hier drei) können paarweise zugeordnet werden, da sich jeder Wert einer Variable für jedes Tool auf das gleiche Liedblatt bezieht. Beispielsweise hat man eine Stichprobe für CER-Werte für das Tool ABBY, Omnipage und Acrobat X Pro und jeder CER-Wert der jeweiligen Stichprobe ist zugehörig zu den CER-Werten der anderen Stichproben, da sie sich auf das gleiche Liedblatt beziehen. Das Tool ist folglich ein Innersubjekt-Faktor. Zu allen Liedblätter werden mehrfach abhängige Variablen erhoben, für jedes Tool einmal. Es ist kein Zwischensubjekt-Faktor, der die Stichprobe nach einer Kategorie trennt. Vorteil dieses Messwiederholungsdesigns ist die „Kontrolle von interindividuellen Unterschieden [...] zwischen den Messungen“ (Leonhart, 2010, S. 179). Es handelt sich auch um das teststärkere Vorgehen, da die Fehlervarianz verringert wird und Leonhart (2010, S. 179) legt nahe dieses Design wenn möglich auch zu wählen.

Wie Kanungo et al. (1998) korrekt feststellen, wird häufig in der bisherigen OCR-Literatur fälschlicherweise angenommen, dass die Stichproben unabhängig voneinander sind. Kanungo et al. (1998) treten auch für die Verwendung des korrekten gepaarten Modells ein, da dies der richtigen statistischen Interpretation entspricht. Auch weisen sie, mit einer Beispielevaluation, einige Vorteile in der Aussagekraft nach. Die Nutzung eines gepaarten Modells führt zur Verwendung anderer statistischer Signifikanztests wie bei unabhängigen Stichproben.

Leonhart (2010) empfiehlt als Signifikanztest für mindestens intervallskalierte Merkmale aus mehr als zwei verbundenen Stichproben die Varianzanalyse mit Messwiederholung. Als Signifikanzniveau wird $p < 0.05$ gewählt.

3.5.3 Voraussetzungen

Zuerst wird an dieser Stelle noch auf die Voraussetzungen zur Durchführung einer Varianzanalyse mit Messwiederholung eingegangen. Nach Leonhart (2013, S. 486) müssen die Variablen mindestens intervallskaliert und normalverteilt sein. Die Bedingung der Skalierung wird erfüllt. Normalverteilung kann über den Test von Shapiro-Wilk nicht nachgewiesen werden (siehe Anhang: Explorative Datenanalyse für jedes Tool). Normalverteilung liegt demnach nicht vor. Nach Lüpsen (2015, S. 21) kann diese Voraussetzung bei ausreichend großer Stichprobengröße ($N > 50$), gemäß dem zentralen Grenzwertsatz, vernachlässigt werden. Auch trägt die gleich große Stichprobengröße dazu bei, dass sich diese Verletzung nicht negativ auf den Test auswirkt (Lüpsen, 2015, S. 21). Generell ist bei der Varianzanalyse die Sphärizitätsannahme deutlich bedeutender um einen Fehler zu vermeiden (Leonhart, 2013, S. 486). Auf dies wird im Weiteren immer explizit eingegangen. Da die Abweichungen von der Normalverteilung jedoch stark sind, wird hier zusätzlich zum Ergebnis der Varianzanalyse das Ergebnis des Friedman-Test für ordinalskalierte Variablen bei mehr als zwei verbundene Stichproben (Leonhart, 2010, S. 177) angegeben. Dabei werden die Daten implizit in eine rang-basierte Ordinalskalierung übertragen um den Test möglich zu machen. Der Test ist schwächer und konservativer, jedoch non-parametrisch und demnach muss keine Normalverteilung vorausgesetzt werden.

3.5.4 Correct In Percent – Beispielhafter Ablauf

3.5.4.1 Signifikanztests – Correct In Percent

Das komplette Verfahren für die Varianzanalyse findet schrittweise statt und wird am folgenden Beispiel für die Variable Correct In Percent erläutert. Zunächst sei hier noch einmal die deskriptive Statistik für die Variable in Bezug auf die drei Tools angegeben. Diese ist notwendig um den Signifikanztest später zu interpretieren.

Tabelle 6: Deskriptive Statistik – Correct In Percent

Deskriptive Statistiken			
	Mittelwert	Standardabweichung	H
Correct In Percent ABBY	79,5895	12,21101	102
Correct In Percent Omnipage	55,7934	27,30632	102
Correct In Percent Acrobat	26,3960	28,29132	102

Die Messwiederholungsfaktoren (Correct In Percent ABBY, Correct In Percent Omnipage, Correct in Percent Acrobat) werden zu einem Messwiederholungsfaktor „Tool“ zusammengefasst.

Als nächstes wird die Sphärizität mit Hilfe des Mauchley-Test überprüft. Dies ist eine sehr wichtige Voraussetzung für die Durchführung der Varianzanalyse, da sonst eine erhöhte Gefahr für einen α -Fehler (Leonhart, 2013, S. 487) besteht. Dabei wird überprüft, ob die Varianz der Differenzwerte innerhalb des Messwiederholungsdesigns über alle Gruppen hinweg gleich ist. Dieser Test muss bei mehr als zwei verbundenen Stichproben durchgeführt werden.

Tabelle 7: Mauchly-Test auf Sphärizität – Correct In Percent**Mauchly-Test auf Sphärizität^a**

Maß: MEASURE_1

Innersubjekt- effekt	Mauchly- W	Näherungs- weise Chi- Quadrat	df	Sig.	Epsilon ^b		
					Green- house-Geis- ser	Huynh-Feldt (HF)	Unter- grenze
Tool	,979	2,096	2	,351	,980	,999	,500

Testet die Nullhypothese, dass die Fehlerkovarianzmatrix der orthonormalisierten transformierten abhängigen Variablen proportional zu einer Identitätsmatrix ist.

a. Design: Konstanter Term

Innersubjekt-Design: Tool

b. Kann für die Anpassung der Freiheitsgrade für die gemittelten Tests auf Signifikanz verwendet werden. Korrigierte Tests werden in der Tabelle 'Tests der Innersubjekteffekte' angezeigt.

Der Test weist kein signifikantes Ergebnis auf (Sig > 0.05). Folglich kann Sphärizität angenommen werden (Rasch, Frieze, Hofmann & Naumann, 2006).

Nun kann man den tatsächlichen Signifikanztest und die Effektstärke für den Faktor Tool in Form des Ergebnisses der einfaktoriellen Varianzanalyse mit Messwiederholung betrachten.

Tabelle 8: Varianzanalyse mit Messwiederholung – Correct In Percent**Tests der Innersubjekteffekte**

Maß: MEASURE_1

Quelle		Typ III Quadrat- summe	df	Quadrati- scher Mittel- wert	F	Sig.	Partielles Eta hoch zwei
Tool	Angenommene Sphä- rizität	144840,347	2	72420,174	149,955	,000	,598
	Greenhouse-Geisser	144840,347	1,959	73922,252	149,955	,000	,598
	Huynh-Feldt (HF)	144840,347	1,998	72503,079	149,955	,000	,598
	Untergrenze	144840,347	1,000	144840,347	149,955	,000	,598
Fehler (Tool)	Angenommene Sphä- rizität	97555,040	202	482,946			
	Greenhouse-Geisser	97555,040	197,895	492,963			
	Huynh-Feldt (HF)	97555,040	201,769	483,499			
	Untergrenze	97555,040	101,000	965,891			

Da Sphärizität angenommen werden kann, betrachtet man die erste Zeile. Hier steht die Wahrscheinlichkeit des F-Werts ($F=149,955$) unter der Nullhypothese (Das Tool hat keinen Einfluss). Die Wahrscheinlichkeit ist mit $p < 0.0001$ deutlich unter dem Signifikanzniveau $\alpha = 0.05$ und damit hochsignifikant. Das partielle Eta-Quadrat ist mit 0.598 weit über der von Bortz (2005, S. 259) vorgeschlagenen Grenze für einen großen Effekt, nämlich 0.14, was also auch die hohe Signifikanz erklärt. Sollte die Sphärizität nicht vorliegen, wird empfohlen die Korrektur nach Greenhouse-Geisser durchzuführen (Leonhart, 2010, S. 182). Ferner kann man feststellen, dass dieser Unterschied linear ist:

Tabelle 9: Zusammenhangstyp – Correct In Percent

Tests der Innersubjektkontraste

Maß: MEASURE_1

Quelle	Tool	Typ III Quadrat- summe	df	Quadrati- scher Mit- telwert	F	Sig.	Partielles Eta hoch zwei	Dezentr. Parameter	Beobach- tete Trenn- schärfe ^a
Tool	Linear	144306,9 66	1	144306,96 6	273,18 8	,000	,730	273,188	1,000
	Quadra- tisch	533,381	1	533,381	1,219	,272	,012	1,219	,194
Fehler (Tool)	Linear	53351,51 1	101	528,233					
	Quadra- tisch	44203,52 8	101	437,659					

a. Berechnet mit $\alpha = ,05$

Wie bereits oben erwähnt wird hier noch das Ergebnis des Friedman-Tests angegeben um etwaige Fehleinschätzungen aufgrund der fehlenden Normalverteilung zu korrigieren. Hierbei werden die Daten in ordinalskalierte Ränge umgewandelt:

Tabelle 10: Deskriptive Statistik – Rangtransformation Correct In Percent

Ränge	
	Mittlerer Rang
Correct In Percent ABBY	2,77
Correct In Percent Omnipage	1,94
Correct In Percent Acrobat	1,29

Insgesamt zeigt sich jedoch auch hier die gleiche Tendenz.

Tabelle 11: Friedman-Test – Correct In Percent

Teststatistiken ^a	
H	102
Chi-Quadrat	113,970
df	2
Asymp. Sig.	,000

a. Friedman-Test

Auch hier kann man ein hochsignifikantes Ergebnis ($p < 0.0001$) feststellen.

Grundsätzlich kann man also die H1 (es gibt einen Unterschied in der Performanz der OCR-Tools ABBY, Omnipage und Adobe Acrobat X Pro für den Testkorpus) in Bezug auf die Erkennungsrate annehmen. Auf Basis des jetzigen Ergebnis kann man jedoch nicht aussagen in Bezug auf welche Beziehung der Unterschied signifikant ist, also ABBY vs. Omnipage, ABBY vs. Acrobat X Pro, Omnipage vs Acrobat X Pro. Hierzu muss man die Haupteffekte über eine Post-Hoc-Analyse miteinander vergleichen. Zur Korrektur der Kumulierung des α -Niveaus bei multiplen Vergleichen wird das Konfidenzintervall mittels Bonferroni-Korrektur (Rasch et al., 2010, S. 3) angepasst.

Tabelle 12: Paarweise Signifikanztests – Correct In Percent

Paarweise Vergleiche						
Maß: MEASURE_1						
(I) Tool	(J) Tool	Mittelwertdifferenz (I-J)	Standardfehler	Sig. ^b	95 % Konfidenzintervall für Differenz ^b	
					Untergrenze	Obergrenze
1	2	23,796*	2,851	,000	16,855	30,737
	3	53,193*	3,218	,000	45,359	61,028
2	1	-23,796*	2,851	,000	-30,737	-16,855
	3	29,397*	3,150	,000	21,729	37,066
3	1	-53,193*	3,218	,000	-61,028	-45,359
	2	-29,397*	3,150	,000	-37,066	-21,729

Basierend auf geschätzten Randmitteln

*. die Mittelwertdifferenz ist auf der Stufe ,05 signifikant.

b. Anpassung für Mehrfachvergleiche: Bonferroni.

Hier erkennt man, dass der Unterschied tatsächlich für jede Kombination von Vergleich hochsignifikant ist ($p < 0.0001$). Tool 1 ist dabei ABBY, Tool 2 Omnipage und Tool 3

Acrobat X Pro. Die erste Zeile 1 vs. 2 steht für die H1.1, Zeile 2, 1 vs. 3 für H1.2, Zeile 4, 2 vs. 3 für H1.3. Die Durchführung des Friedman-Tests für alle drei Kombinationen führt zum selben Ergebnis.

Tabelle 13: Deskriptive Statistik – Rangtransformation Paarweise I

Ränge	
	Mittlerer Rang
Correct In Percent ABBY	1,86
Correct In Percent Omni-page	1,14

Tabelle 14: Friedman-Test – Paarweise I

Teststatistiken ^a	
H	102
Chi-Quadrat	54,760
df	1
Asymp. Sig.	,000

a. Friedman-Test

Tabelle 15: Deskriptive Statistik – Rangtransformation Paarweise II

Ränge	
	Mittlerer Rang
Correct In Percent ABBY	1,91
Correct In Percent Acrobat	1,09

Tabelle 16: Friedman-Test – Paarweise II

Teststatistiken ^a	
H	102
Chi-Quadrat	69,176
df	1
Asymp. Sig.	,000

a. Friedman-Test

Tabelle 17: Deskriptive Statistik – Rangtransformation Paarweise III

Ränge	
	Mittlerer Rang
Correct In Percent Acrobat	1,20
Correct In Percent Omnipage	1,80

Tabelle 18: Friedman-Test – Paarweise III

Teststatistiken ^a	
H	102
Chi-Quadrat	36,842
df	1
Asymp. Sig.	,000

a. Friedman-Test

Durch Interpretation der deskriptiven Daten kann man auch die Richtung des Vergleichs feststellen. ABBY ist bezüglich der Variable Correct in Percent, also der Genauigkeit, signifikant besser als Omnipage und Acrobat X Pro. Omnipage ist signifikant besser als Acrobat X Pro. Die H1.1, H1.2, H1.3 können für die untersuchte Variable auch angenommen werden.

Für die weiteren Performanzvariablen wird das statistische Vorgehen nicht mehr vollständig geschildert, die Ergebnisse werden lediglich tabellarisch aufbereitet und erläutert. Für den Wert Correct In Percent sei diese Zusammenfassung hier auch noch angegeben:

Tabelle 19: Signifikanztests Zusammenfassung – Correct In Percent

Sphärizität	F-Wert	Signifikanz	Partielles Eta-Quadrat	Signifikanz Friedman
p = 0.351	149,955	p < 0.0001	0.598	p < 0.0001

Ist die Signifikanz für Sphärizität kleiner 0.05 werden die Ergebnisse der Greenhouse-Geisser-Korrektur angegeben. Der F-Wert gibt die Stärke des Effekts an (Tool). Die Signifikanz gibt an ob die H1 für diese Variable angenommen wird. Aufgrund der Verletzung der Normalverteilung wird die Signifikanz nach Friedman zur Kontrolle mitangegeben. Das partielle Eta-Quadrat ist ein weiteres Maß für die Stärke des Effekts und gibt den Anteil der aufgeklärten Varianz über die Gesamtvarianz an. Ein Wert größer 0.14 weist auf einen starken Effekt hin. Wenn der Signifikanztest hier also einen signifikanten Unterschied aufzeigt, werden die Ergebnisse der paarweisen Vergleiche angegeben:

Tabelle 20: Paarweise Signifikanztests Zusammenfassung – Correct In Percent

Vergleich	Signifikanz	Signifikanz Friedman
ABBY vs. Omnipage	$p < 0.0001$	$p < 0.0001$
ABBY vs. Acrobat X Pro	$p < 0.0001$	$p < 0.0001$
Omnipage vs Acrobat X Pro	$p < 0.0001$	$p < 0.0001$

Ist die Signifikanz, wie hier, kleiner 0.05 können H1.1, H1.2 und H1.3 angenommen werden. Es muss durch die Interpretation der deskriptiven Daten noch festgestellt werden in welche Richtung der Unterschied geht, also welches Tool besser ist als das andere.

Alle Auswertungen für jede Variable finden sich komplett im Anhang als SPSS-Vierwer-Dateien mit dem Namen Inferenzstatistik-[Variable].

3.5.4.2 Boxplot-Grafik – Correct In Percent

Die Ergebnisse lassen sich mit einer deskriptiven Visualisierung über Boxplots noch verdeutlichen:

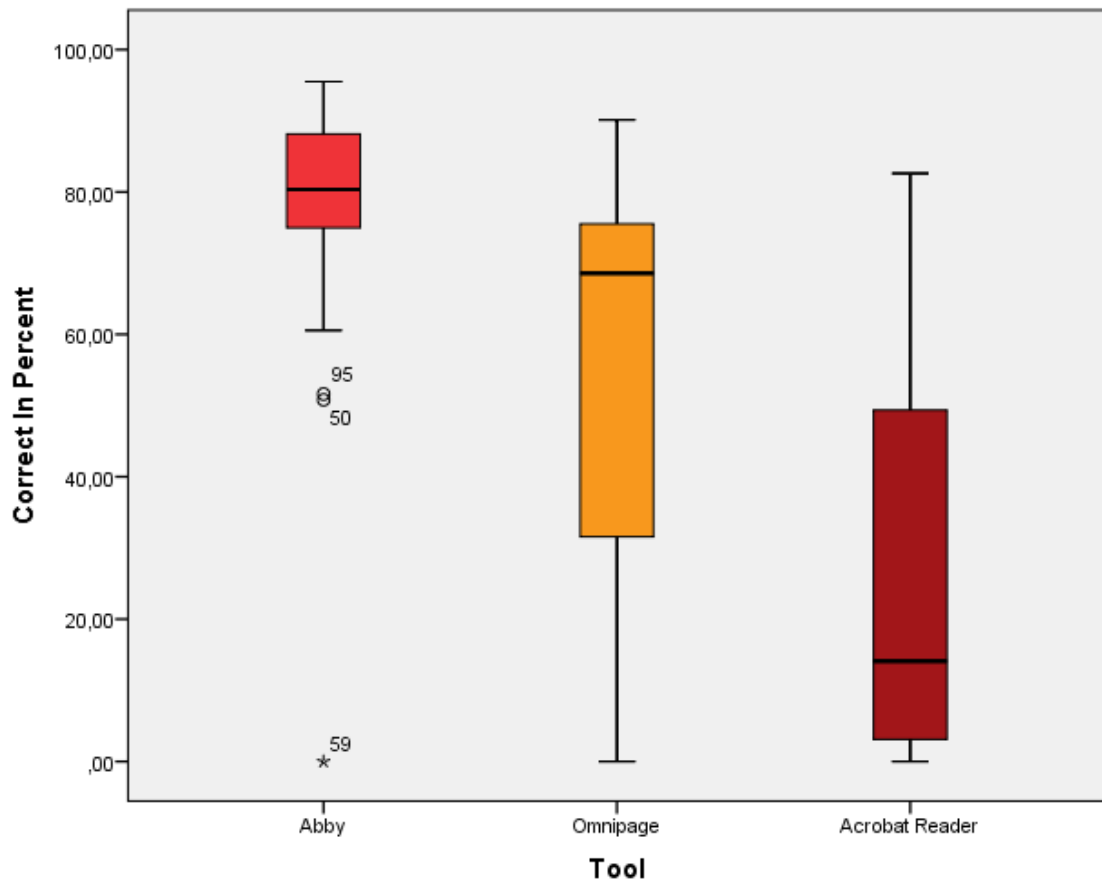
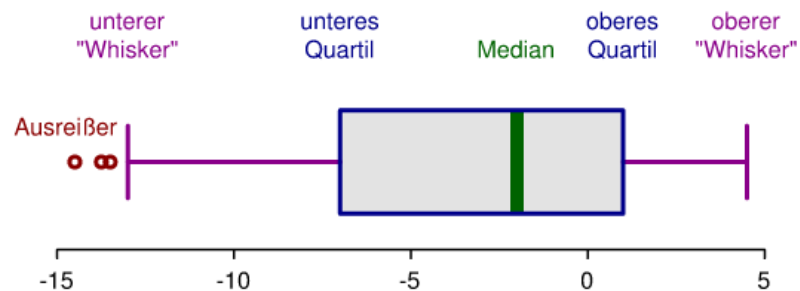


Abbildung 36: Boxplot-Grafik – Correct In Percent

Die Ergebnisse der Signifikanztests lassen sich hier visuell interpretieren. Folgende Grafik hilft bei dem Verständnis eines Boxplots:



8

Abbildung 37: Boxplot Beispiel

Der Median ist als Strich in der Box markiert. Man muss anmerken, dass sich der Median hier sehr stark vom Mittelwert unterscheiden kann. Der untere „Whisker“ und obere „Whisker“ sind das Maximum, respektive Minimum, wobei dabei Ausreißer und Extremwerte, welche die tatsächlichen Minima oder Maxima bilden, im Boxplot als extra

⁸ Quelle :https://upload.wikimedia.org/wikipedia/commons/b/b1/Elements_of_a_boxplot.svg

Punkte erscheinen. Ausreißer liegen mehr als das 1,5 – fache vom Quartilsabstand entfernt (Kreisform). Extremwerte (Sternchenform) mehr als das 3 – fache entfernt. Das untere Quartil enthält 25% der Datenwerte, die kleiner als der Median sind. Das obere Quartil 25% der Datenwerte, die größer als der Median sind. Dementsprechend befinden sich von einem Boxende bis zum Whisker, jeweils die restlichen 25%.

Hier sieht man also, dass ABBY von allen drei Tools mit einem Median von 80% die beste Leistung für die Erkennungsrate bringt, und vor allen Dingen eine deutlich geringere Streuung aufweist, folglich stabiler in der Erkennung ist. Jedoch gibt es auch bei ABBY einige Ausreißer nach unten. Der Median von Omnipage liegt deutlich über dem Mittelwert von 55%, nämlich bei 68%. Das oberste Quartil erreicht jedoch nur 90%. Das heißt alle Werte befinden sich unter 90% und damit dem notwendigen Zielwert für brauchbare OCR-Erkennung (Holley, 2009). Acrobat X Pro hat seinen Median nur bei 15%. Das heißt die Hälfte aller Blätter wird unter diesem Wert erkannt. Die Streuung bei Acrobat ist sehr groß. Es werden maximal Werte bis 80% erreicht.

3.5.5 Precision

3.5.5.1 Signifikanztests – Precision

Wie bereits in der deskriptiven Statistik gezeigt, lassen sich über die Precision nur geringfügig andere Ergebnisse zeigen als durch die Accuracy. Hier sei nochmal ein Ausschnitt der Daten wiederholt:

Tabelle 21: Deskriptive Statistik – Precision

Deskriptive Statistiken			
	Mittelwert	Standardabweichung	H
Precision ABBY	73,5722	12,92064	102
Precision Omnipage	52,0811	25,91975	102
Precision Acrobat	21,9626	24,02734	102

Man sieht also, dass eine ähnliche Tendenz wie für Correct In Percent besteht. Die Inferenzstatistik kann dies nachweisen:

Tabelle 22: Signifikanztests Zusammenfassung – Correct In Percent

Sphärizität	F-Wert	Signifikanz	Partielles Eta-Quadrat	Signifikanz Friedman
p = 0.691	1677.008	p < 0.0001	0.623	p < 0.0001

Es besteht ein signifikanter Unterschied in der Performanz, gemessen an der Precision bezüglich der drei Tools. Dies kann sowohl mit dem parametrischen als auch dem non-parametrischen Test gezeigt werden. Der Unterschied ist hochsignifikant. Auch die gepaarten Vergleiche sind hoch signifikant:

Tabelle 23: Paarweise Signifikanztests Zusammenfassung – Precision

Vergleich	Signifikanz	Signifikanz Friedman
ABBY vs. Omnipage	p < 0.0001	p < 0.0001
ABBY vs. Acrobat X Pro	p < 0.0001	p < 0.0001
Omnipage vs Acrobat X Pro	p < 0.0001	p < 0.0001

Die H1, H1.1, H1.2 und H1.3 können alle angenommen werden. ABBY ist damit nicht nur in Bezug auf den Recall (Correct In Percent), also dem generellen Anteil der erkannten Zeichen aus dem Grounded Truth am besten, sondern auch in Bezug auf die Precision. Von allen drei Tools muss im OCR-Output bei ABBY am wenigsten ausgebessert werden, seien es Fehler oder Noise.

3.5.5.2 Boxplots-Grafik – Precision

Zur Vervollständigung sei dieser Sachverhalt an der folgenden Boxplot-Grafik aufgezeigt:

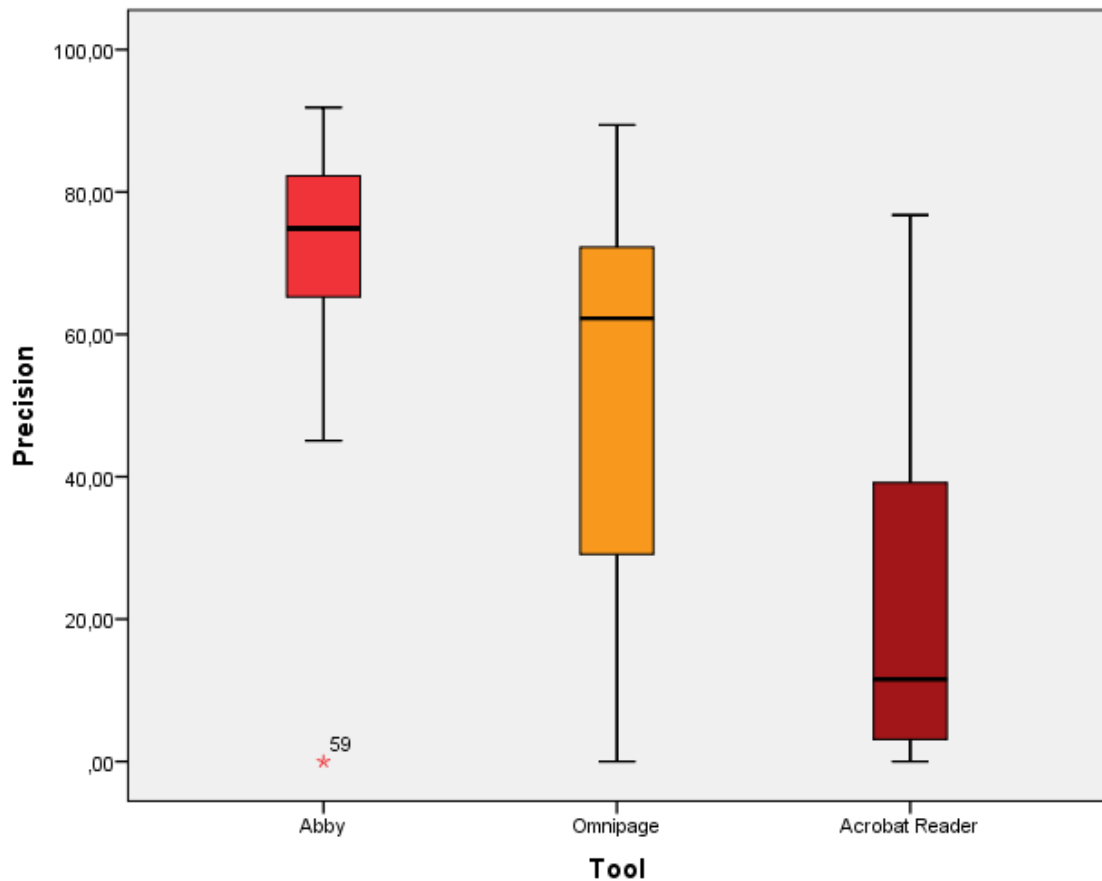


Abbildung 38: Boxplot-Grafik – Precision

Das Bild ist ähnlich zu oben (Correct In Percent), lediglich bei allen drei Tools geringfügig schlechter. Streuung, Mittelwerte und Median sind äquivalent. ABBY ist auch nach visueller Interpretation das Tool mit der stabilsten Leistung.

3.5.6 Spurious In Percent

3.5.6.1 Signifikanztests – Spurious In Percent

Im Bereich der Spurious-Werte, also ‚der Noise‘ im OCR-Output liegen folgende Mittelwerte vor:

Tabelle 24: Deskriptive Statistik – Spurious In Percent

Deskriptive Statistiken			
	Mittelwert	Standardabweichung	H
Spurious In Percent ABBY	8,6452	6,03799	102
Spurious In Percent Omnipage	8,0192	19,46479	102
Spurious In Percent Acrobat	19,0471	59,47921	102

Bezüglich der Inferenzstatistik zur H1 stellt man folgende Werte fest:

Tabelle 25: Signifikanztests Zusammenfassung – Spurious In Percent

Sphärizität	F-Wert	Signifikanz	Partielles Eta-Quadrat	Signifikanz Friedman
$p < 0.0001$	2.909	$p = 0.085$	0.028	$p < 0.0001$

Sphärizität liegt nicht vor; es wird auf das Korrekturverfahren von Greenhouse-Grasser zurückgegriffen. Der p-Wert ist jedoch größer 0.05. Allerdings weist der non-parametrische Test über Friedman eine hohe Signifikanz auf. Wie weiter oben beschrieben ist die Aussagekraft des Friedman-Tests geringer. Die H1 bezüglich Spurious in Percent kann also nur bedingt angenommen werden. Für die paarweisen Vergleiche erhält man folgende Ergebnisse:

Tabelle 26: Paarweise Signifikanztests Zusammenfassung – Spurious In Percent

Vergleich	Signifikanz	Signifikanz Friedman
ABBY vs. Omnipage	$p = 1.000$	$p < 0.0001$
ABBY vs. Acrobat X Pro	$p = 0.262$	$p < 0.0001$
Omnipage vs Acrobat X Pro	$p = 0.239$	$p = 0.001$

Auch hier kann über die einfaktorielle Varianzanalyse mit Messwiederholung kein signifikanter Unterschied nachgewiesen werden. Über den Friedman-Test erhält man jedoch schon signifikante Ergebnisse. Die visuelle Interpretation lässt auch den Schluss zu, dass entscheidende Unterschiede vorliegen, vor allem in Bezug auf das Tool ABBY. H1.1, H1.2, H1.3 kann also für die Variable Spurious In Percent bedingt angenommen werden.

3.5.6.2 Boxplots-Grafik – Spurious In Percent

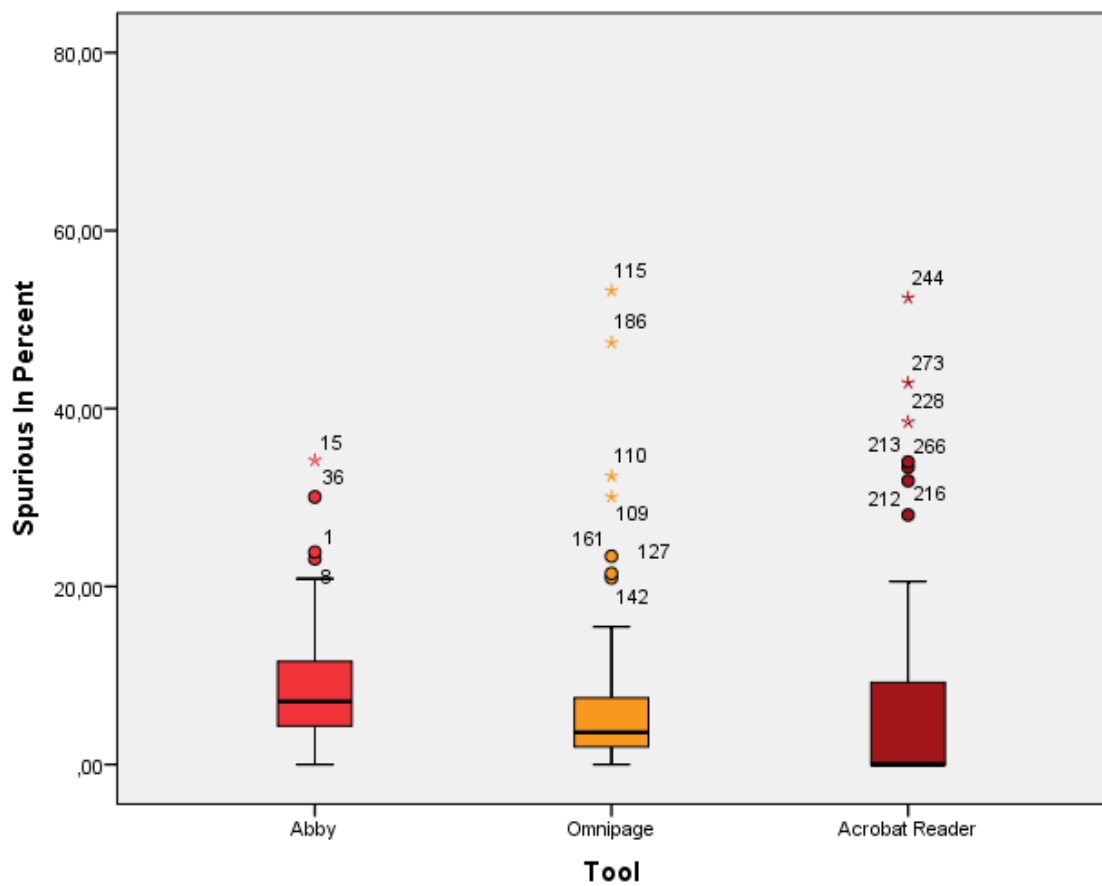


Abbildung 39: Spurious In Percent

Der Wertebereich der Boxplot-Grafik wurde zur besseren Übersichtlichkeit auf 80% verringert. Tatsächlich haben sowohl Acrobat X Pro als auch Omnipage noch vereinzelte Extremwerte außerhalb dieses Wertebereichs, im Fall von Omnipage bis 182% und für Acrobat bis 400%. Bis auf die Ausreißer sind die Verhältnisse folglich ähnlich, jedoch haben sowohl Omnipage als auch Acrobat Ausgaben mit sehr großem Noise-Anteil produziert. ABBY hat nur eine geringfügige Anzahl von Ausreißern.

3.5.7 Confused In Percent

3.5.7.1 Signifikanztests – Confused In Percent

Für die Variable Confused In Percent, also wie viele Zeichen des Grounded-Truth falsch ermittelt wurden, zeigt sich eine etwas andere Tendenz wie bisher:

Tabelle 27: Deskriptive Statistik – Confused In Percent

Deskriptive Statistiken			
	Mittelwert	Standardabweichung	H
Confused In Percent ABBY	15,5821	7,94531	102
Confused In Percent Omnipage	19,9421	14,81493	102
Confused In Percent Acrobat	15,7189	16,59953	102

Tabelle 28: Signifikanztests Zusammenfassung – Confused In Percent

Sphärizität	F-Wert	Signifikanz	Partielles Eta-Quadrat	Signifikanz Friedman
p = 0.01	3.506	p = 0.036	0.034	p = 0.011

Auch hier wird die Bedingung der Sphärizität nicht erfüllt. Es wird auf die Greenhouse-Grasser-Korrektur ausgewichen und diese Werte hier angegeben. Bezüglich des Signifikanzniveaus von 0.05 kann die H1 angenommen werden. Über die paarweisen Vergleiche kann man den genauen Hauptunterschied identifizieren:

Tabelle 29: Paarweise Signifikanztests Zusammenfassung – Confused In Percent

Vergleich	Signifikanz	Signifikanz Friedman
ABBY vs. Omnipage	p = 0.022	p = 0.028
ABBY vs. Acrobat X Pro	p = 1.000	p = 0.037
Omnipage vs Acrobat X Pro	p = 0.135	p = 0.035

Tatsächlich ist mit der Varianzanalyse nur die H1.1 nachweisbar. Also der Unterschied bezüglich der Confused-In-Percent-Werte ist für ABBY und Omnipage signifikant. Für die anderen Vergleiche nicht. Lediglich der Test nach Friedman weist einen signifikanten jedoch sehr moderaten Unterschied nach. Auch visuell lässt sich erkennen, dass der

Unterschied zwischen ABBY und Acrobat X Pro sowie Omnipage und Acrobat X Pro recht gering ist, weswegen die H1.2 und H1.3 verworfen werden. Ohnehin muss die H1.3 verworfen werden, da die Richtung des Unterschieds nicht stimmen würde.

3.5.7.2 Boxplots-Grafik – Confused In Percent

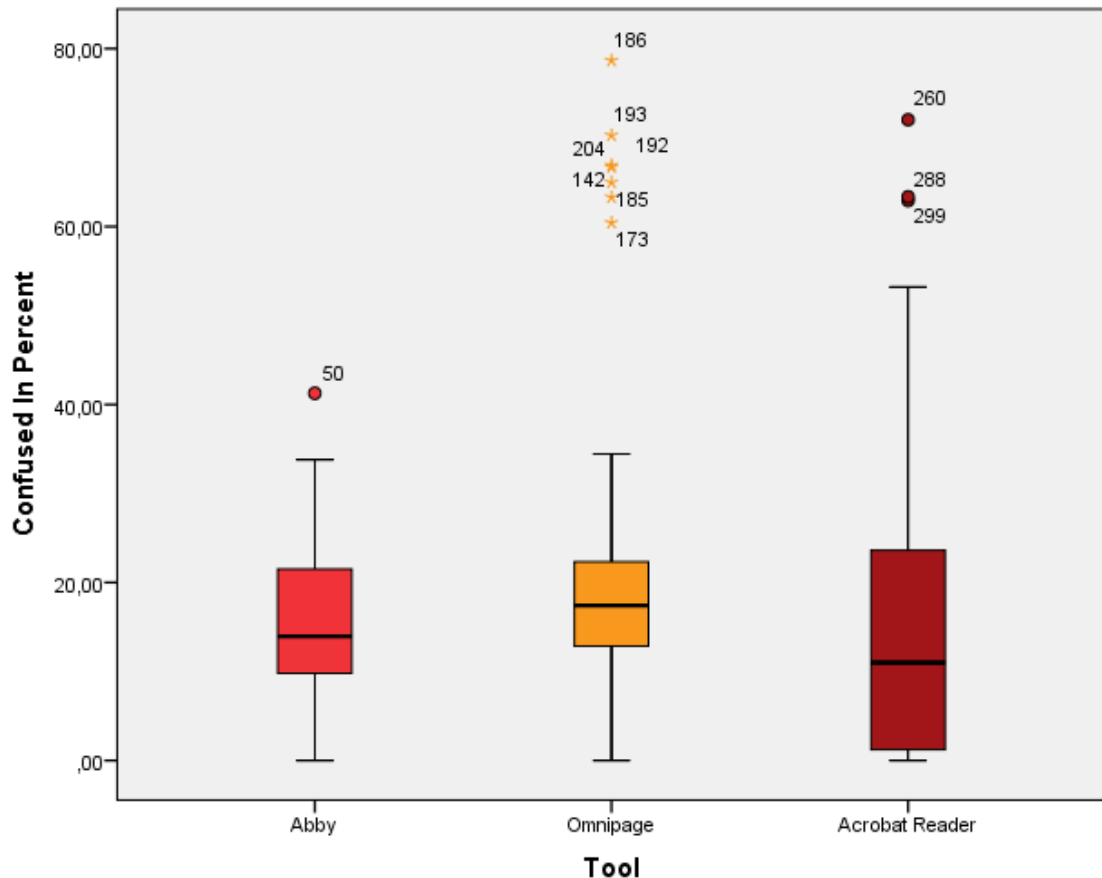


Abbildung 40: Boxplot-Grafik – Confused In Percent

Auch für Confused In Percent erkennt man, dass für ABBY die Streuung deutlich geringer ist als bei den anderen Tools, wie man auch aus der deskriptiven Statistik an der Standardabweichung erkennen kann. Obwohl also alle drei Tools einen naheliegenden Median haben (für ABBY 14%, Omnipage 17%, Acrobat 10%), ist ABBY, in Bezug auf die untersuchte Variable, konstanter in der Verarbeitung. Omnipage hat einen ähnlichen Boxplot wie ABBY, jedoch vereinzelte Extremwerte. Acrobat hat tatsächlich bei einem größeren Teil von Blättern eine bessere Performanz, die Hälfte wird mit einer Confusion-Rate unter 10% erkannt. Die Streuung nach oben ist jedoch größer als bei den anderen Tools. Bezüglich eines signifikanten Unterschieds gleicht sich dies jedoch wieder aus.

3.5.8 Lost In Percent

3.5.8.1 Signifikanztests – Lost In Percent

Tabelle 30: Deskriptive Statistik – Lost In Percent

Deskriptive Statistiken			
	Mittelwert	Standardabweichung	H
Lost In Percent ABBY	4,8285	11,19534	102
Lost In Percent Omnipage	24,2645	29,52750	102
Lost In Percent Acrobat	57,8852	40,02903	102

Die deskriptiven Daten lassen eine signifikante Tendenz vermuten. Die Ergebnisse der Signifikanztests bestätigen diese Aussage für die Variable Lost In Percent:

Tabelle 31: Signifikanztests Zusammenfassung – Lost In Percent

Sphärizität	F-Wert	Signifikanz	Partielles Eta-Quadrat	Signifikanz Friedman
p < 0.0001	95.689	p < 0.0001	0.486	p < 0.0001

Sphärizität liegt nicht vor. Hier sind die Werte über die Greenhouse-Geisser-Korrektur angegeben. Sowohl F-Wert als auch das partielle Eta-Quadrat belegen einen starken Effekt des Tools auf die Lost-In-Percent-Werte. Dieser Effekt ist hochsignifikant, sowohl für den parametrischen als auch den non-parametrischen Test. Die H1 kann für diese Variable angenommen werden. Die paarweisen Tests zeigen, dass bezüglich Lost In Percent für jeden Vergleich ein signifikanter Unterschied besteht.

Tabelle 32: Paarweise Signifikanztests Zusammenfassung – Lost In Percent

Vergleich	Signifikanz	Signifikanz Friedman
ABBY vs. Omnipage	p < 0.0001	p < 0.0001
ABBY vs. Acrobat X Pro	p < 0.0001	p < 0.0001
Omnipage vs Acrobat X Pro	p < 0.0001	p < 0.0001

Alle Hypothesen können folglich eindeutig angenommen werden. In der Boxplot-Darstellung wird dies besonders klar.

3.5.8.2 Boxplots-Grafik – Lost In Percent

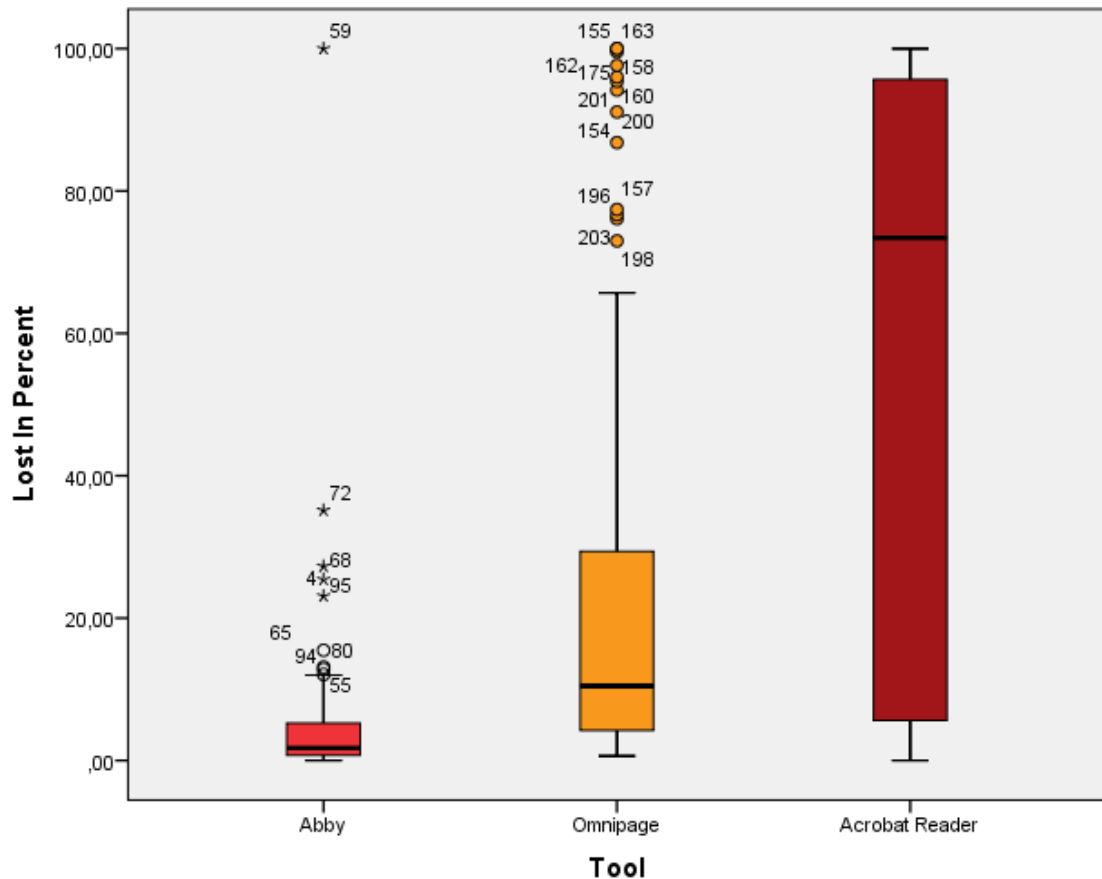


Abbildung 41: Boxplot-Grafik – Lost In Percent

Bis auf vereinzelte Ausreißer und Extremwerte hat ABBY für fast 100% aller Blätter einen Lost-Anteil zwischen 0 und 20%. Omnipage hat eine ähnliche Verteilungsform, und die Hälfte der Blätter weisen einen Lost-Anteil unter 10% auf. Die Streuung nach oben ist jedoch größer und es liegen mehr Ausreißer vor. Acrobat hat einen Median von 73%. Somit werden von mehr als der Hälfte der Liedblätter fast alle Zeichen verloren. Die Streuung ist stark, der Anteil an hohen Verlustraten sehr hoch. Das größte Leistungsdefizit von Acrobat ist tatsächlich in Bezug auf die Lost-Werte zu sehen. Das erklärt auch die verhältnismäßig guten Raten für z.B. die Confused-Werte, da andere Fehlertypen nicht mehr auftreten können wenn der Mehranteil eines Blattes gar nicht erkannt wird. Die Grafik verdeutlicht, dass alle Hypothesen angenommen werden müssen.

3.5.9 Character Error Rate

3.5.9.1 Signifikanztests – CER

Die Variable CER fasst alle Fehler in Bezug zum Grounded-Truth-Text zusammen und ist demnach ähnlich zur Accuracy und Precision eine Hauptmetrik. Das Ergebnis der Hypothesentests ist infolgedessen, insgesamt betrachtet, sehr bedeutsam.

Tabelle 33: Deskriptive Statistik – CER

Deskriptive Statistiken			
	Mittelwert	Standardabweichung	H
CER ABBY	29,0554	14,62504	102
CER Omnipage	52,2256	33,09877	102
CER Acrobat	92,6514	58,58258	102

Tabelle 34: Signifikanztests Zusammenfassung – CER

Sphärizität	F-Wert	Signifikanz	Partielles Eta-Quadrat	Signifikanz Friedman
$p < 0.0001$	69.743	$p < 0.0001$	0.408	$p < 0.0001$

Auch hier liegt keine Sphärizität vor; es wird also so verfahren wie in den obigen Fällen. Der Unterschied zwischen den Tools ist bezüglich des CER hochsignifikant. Auch die F-Werte und das partielle Eta-Quadrat sind überdurchschnittlich groß. Die H1 wurde demnach belegt.

Tabelle 35: Paarweise Signifikanztests Zusammenfassung – CER

Vergleich	Signifikanz	Signifikanz Friedman
ABBY vs. Omnipage	$p < 0.0001$	$p < 0.0001$
ABBY vs. Acrobat X Pro	$p < 0.0001$	$p < 0.0001$
Omnipage vs Acrobat X Pro	$p < 0.0001$	$p < 0.0001$

Die Unterschiede sind signifikant für jeden Vergleich. Ähnlich zur Accuracy und zu Lost In Percent lassen sich die Werte für CER sehr deutlich über den Tool-Effekt belegen. Alle Unterhypothesen H1.1, H1.2, H1.3 werden angenommen. ABBY ist besser als Omnipage und Acrobat X Pro und Omnipage besser als Acrobat X Pro in Bezug auf die Fehlerrate. Dies erkennt man auch in der Boxplot-Grafik.

3.5.9.2 Boxplots-Grafik – CER

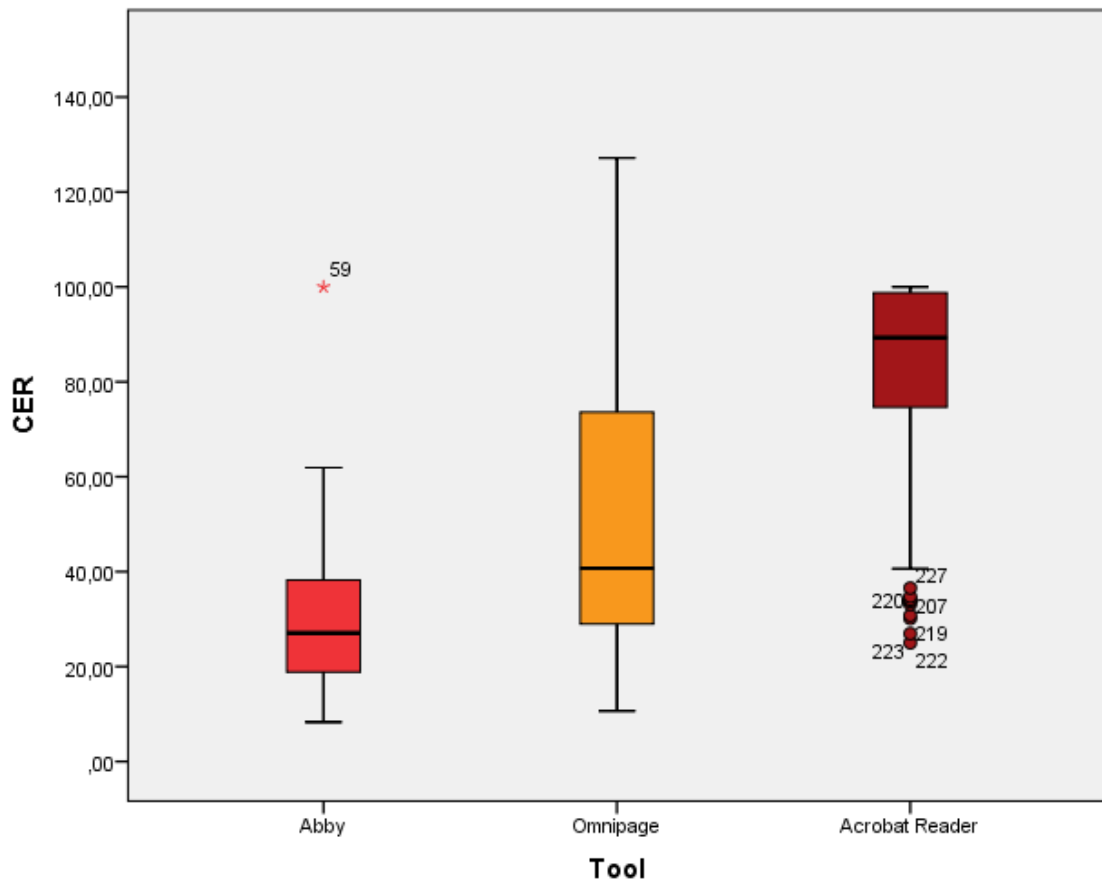


Abbildung 42: Boxplot-Grafik – CER

Der Wertebereich dieser Grafik wurde auch angepasst, da Omnipage und Acrobat Ausreißer und Extremwerte bis 250% bzw. 460% haben.

Auch hier überzeugt ABBY durch eine sehr stabile Verteilung mit einer geringen Standardabweichung und einem Median bei 27%. Der Median von Omnipage liegt bei 40% und auch hier ist wieder vor allem die Streuung nach oben bis zum oberen Quartilsbereich von 130% und dem Ausreißer bis 250% (hier nicht mehr dargestellt) sehr stark. Mit einem Median von 90 Prozent schneidet Acrobat am schlechtesten ab. Einige Ausreißer nach unten stellen Blätter mit einer besseren Fehlerleistung dar. Der untere „Whisker“ befindet sich jedoch bei 40%. Dies ist die Grenze bis zu welcher die Fehlerrate

von 75% aller Blätter bei ABBY verläuft. Insgesamt ist der Unterschied der Tools für die Hauptmetrik CER also sehr deutlich und nachweislich signifikant.

3.5.10 Nachbemerkung – Frakturschrift

Alle statistischen Tests wurden auch hinsichtlich der Frage unternommen, ob die Frakturschrift einen Einfluss auf die Performanz hat. Tatsächlich konnten ambivalente Ergebnisse festgestellt werden. So ist die Performanz von ABBY bzgl. Correct In Percent annähernd gleich im Bereich von Fraktur und ohne Fraktur. Omnipage und Acrobat zeigen jedoch eine deutlich bessere Leistung bzgl. Fraktur. Omnipage eine Steigerung um 15% und Acrobat sogar um 50% auf 75%. Die Erklärung könnte daran liegen, dass die Druckstärke und damit der Kontrast bei der Frakturschrift deutlich stärker ist als bei der Schreibmaschinenschrift welche oft nur sehr schwach gedruckt ist. Die bisherige Analyse des Korpus legt jedoch nahe, dass die Mehrzahl der Liedblattsammlung nicht aus diesem speziellen Frakturblättern besteht. Deswegen wurde diese Bemerkung nicht weiterverfolgt.

3.5.11 Weitere vergleichende Visualisierungen

In diesem Kapitel werden noch einige Grafiken der deskriptiven Statistik gezeigt, die explizit die bisher festgestellten Aussagen visualisieren. Es handelt sich dabei um vergleichende Grafiken. Die oben schon gezeigten Boxplots-Grafiken gehören auch zu diesem Bereich der deskriptiven Statistik.

3.5.11.1 Scatterplots

Ein Scatterplot, oder auch Streudiagramm ist die graphische Darstellung zweier Wertpaare. Ein Liedblatt hat für jedes Tool einen Performanzparameter. Die Leistung zweier Tools kann man demnach mit einem Scatterplot visualisieren. Der x- bzw. y-Wert sind die entsprechenden Werte der zu untersuchenden Variable zu dem Liedblatt. Kanungo et al. (1998) verwenden diese Visualisierung für ihr OCR-Forschungsprojekt.

Folgende Scatterplots sollen noch mal die Einzelhypothesen H1.1, H1.2 und H1.3 grafisch darstellen. Man beschränkt sich dabei auf die Hauptmetriken Correct In Percent, Character Error Rate und die Lost-Rat. Für die Lost-Rate konnte gezeigt werden, dass sie eine entscheidende Rolle bei den Performanzunterschieden spielt. Die Precision ist zwar auch eine Hauptmetrik, es wurde aber schon beschrieben, dass dieses Maß mit Correct In Percent stark korreliert und an dieser Stelle nur redundante Informationen

bringen würde. Nachfolgende Grafik soll die Art und Weise der Interpretation der Scatterplots erklären:

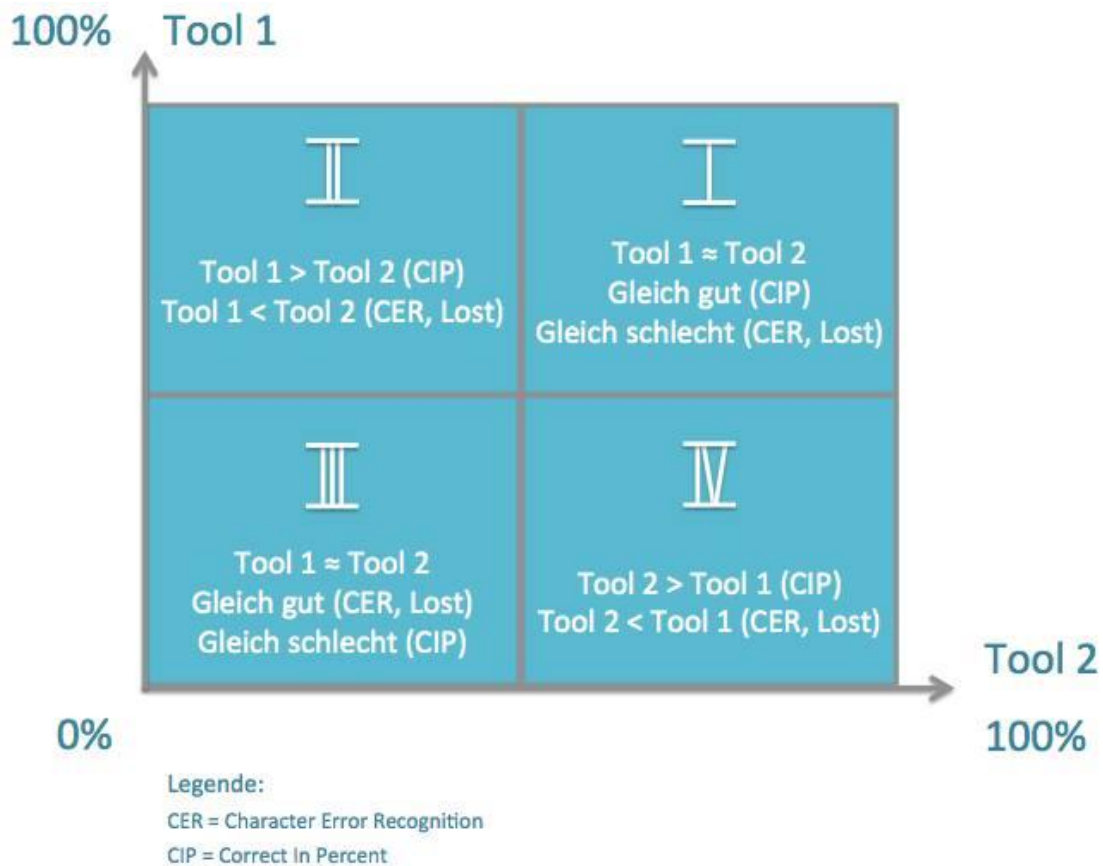


Abbildung 43: Scatterplot Interpretation

Befinden sich Punkte (also Liedblätter) im Quadranten I, ist die Performanz der beiden Tools annähernd gleich gut bei der Metrik Correct In Percent. Für Lost In Percent und CER heißt es, dass die Tools annähernd gleich schlecht sind. Befinden sich Punkte im Quadranten III weist das auch auf gleiche Performanz für diese Blätter hin, diesmal nur umgekehrt. Befinden sich Punkte im Quadranten II heißt das für Correct In Percent, dass für diese Blätter Tool 1 besser ist als Tool 2. Bei CER und Lost in Percent ist Tool 1 schlechter. Für den Quadranten IV gilt ähnliches nur mit umgekehrten Aussagen. Hier ist Tool 2 besser als Tool 1 für die Accuracy und schlechter für CER und Lost in Percent.

Dies sind die Scatterplots für Correct In Percent:

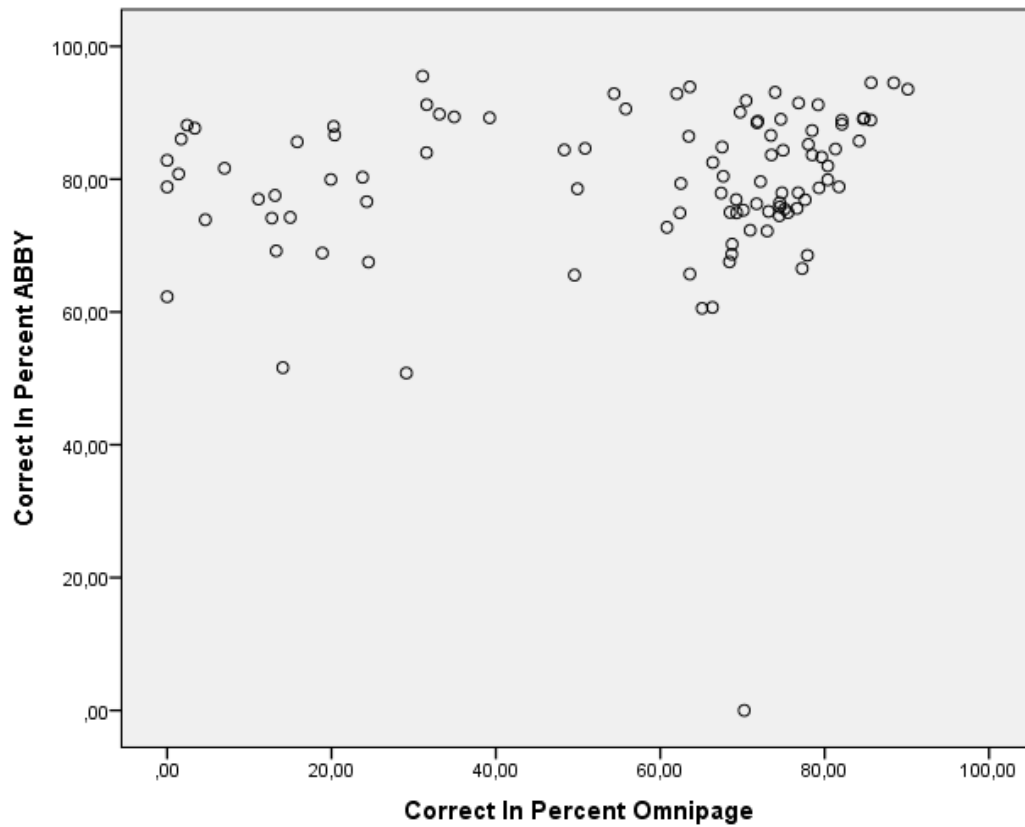


Abbildung 44: Scatterplot – Correct In Percent ABBY-Omnipage

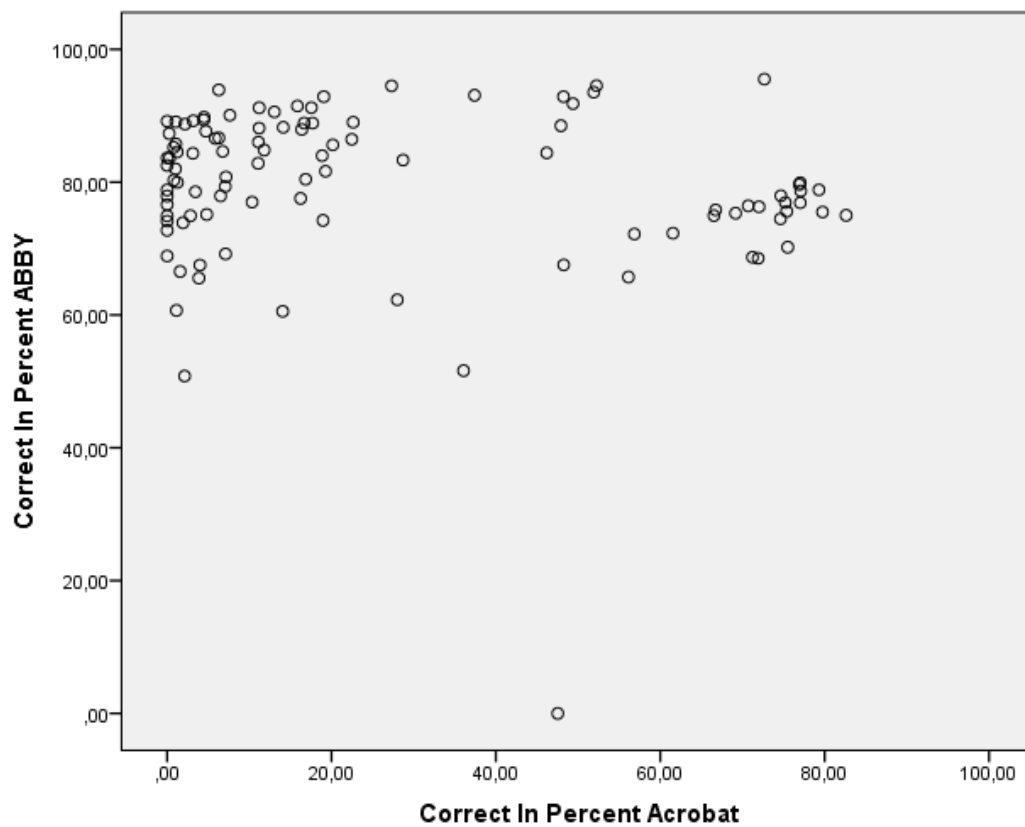


Abbildung 45: Scatterplot – Correct In Percent ABBY/Acrobat

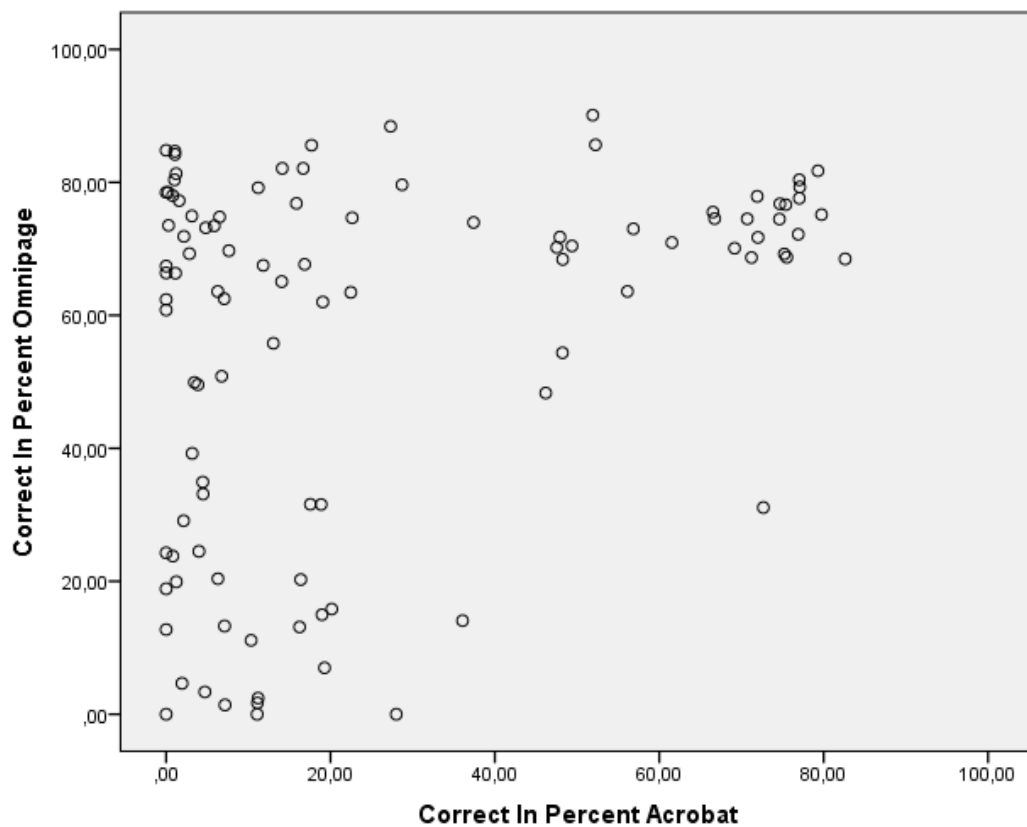


Abbildung 46: Scatterplot – Correct In Percent Omnipage/Acrobat

Bei der ersten Grafik lässt sich erkennen, dass ABBY und Omnipage verhältnismäßig starke Erkennungsraten haben, da ein großer Teil der Blätter im 1. Quadranten sind. Ein weiterer großer Teil ist auch im 2. Quadranten, was die H1.1 wieder bestätigt. Ein ähnliches Bild zeigt sich beim zweiten Scatterplot, nur viel deutlicher. Der Großteil aller Liedblätter ist hier im 2. Quadranten. ABBY hat hier also eine deutlich bessere Leistung als Acrobat X Pro. Bei der Unterscheidung zwischen Omnipage und Acrobat X Pro sieht man die H1.3 auch bestätigt, jedoch gibt es auch zahlreiche Blätter für die beide eine ähnlich schlechte Erkennungsrate haben.

Die Grafiken für Lost In Percent zeichnen ein deutliches Bild:

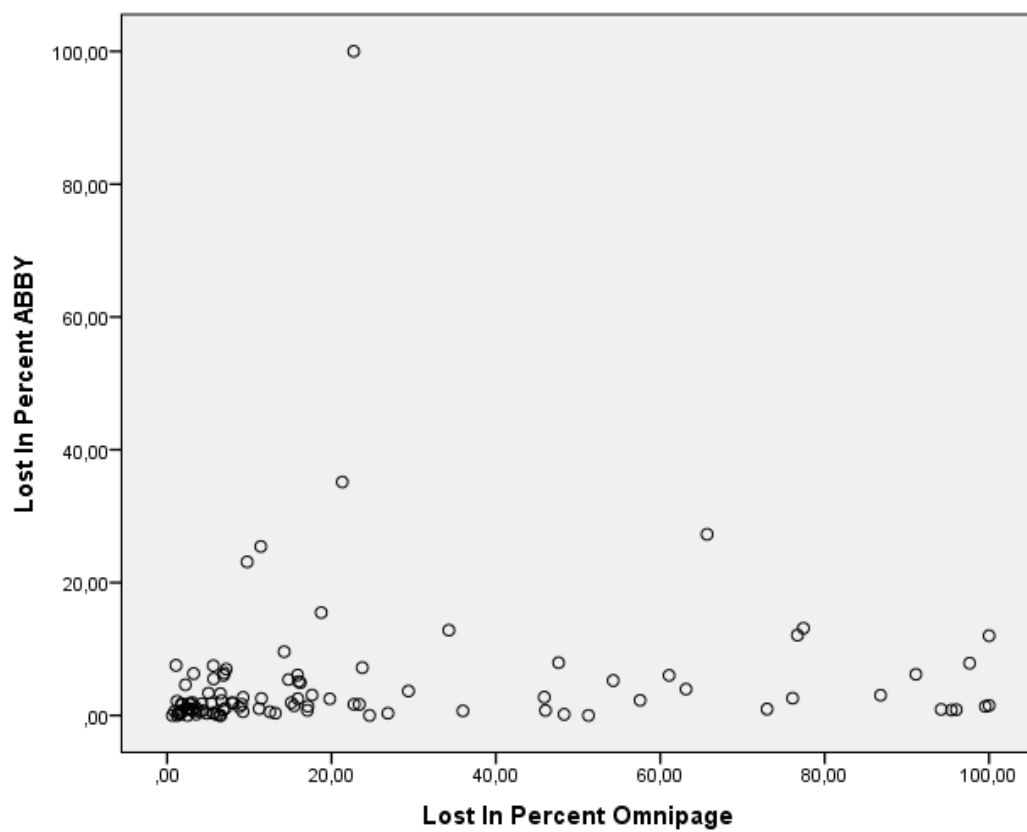


Abbildung 47: Scatterplot – Lost In Percent ABBY-Omnipage

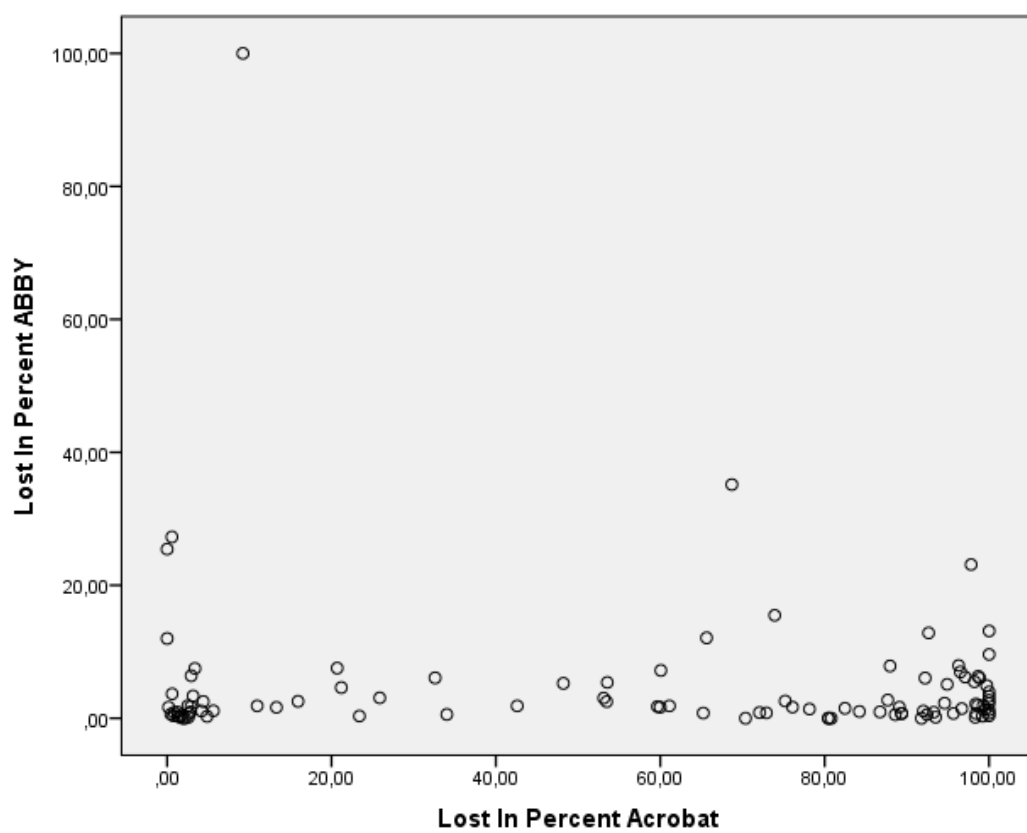


Abbildung 48: Scatterplot – Lost In Percent ABBY/Acrobat

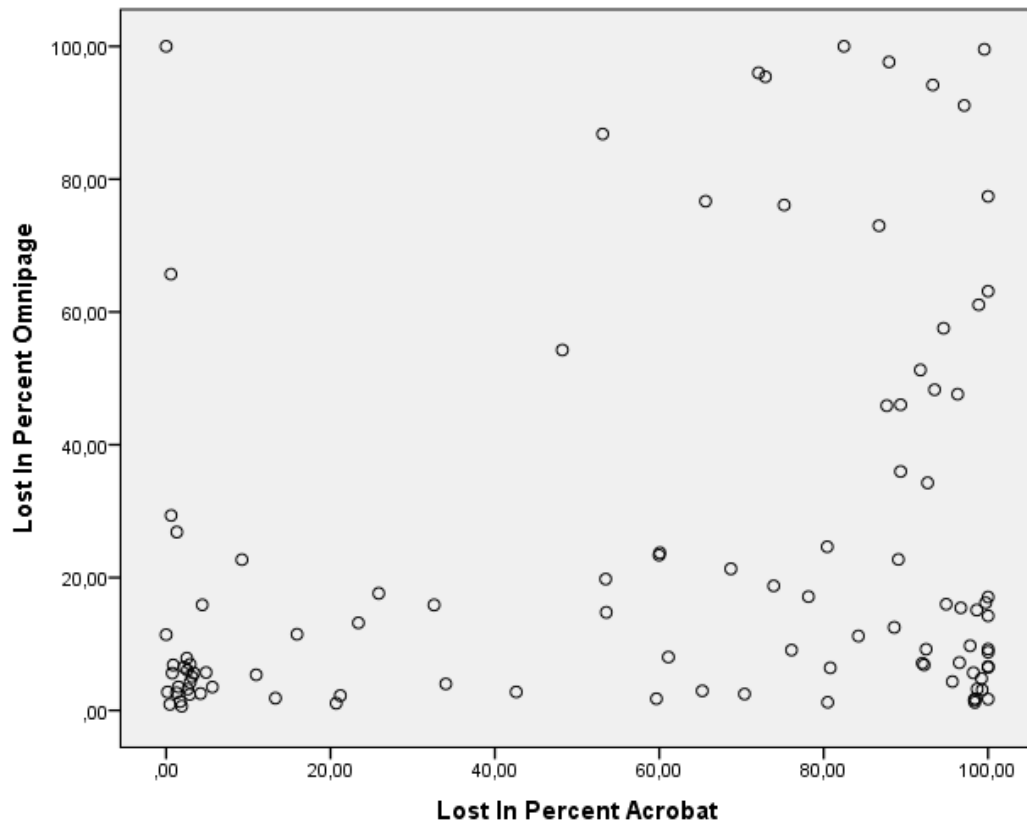


Abbildung 49: Scatterplot – Lost In Percent Omnipage-Acrobat

Auch hier befinden sich die Dokumente beim Vergleich von ABBY und Omnipage hauptsächlich im positiven Quadranten, also hier Quadrant III. Eine kleine Menge in Quadrant IV bestätigt aber die überdurchschnittlich starke Performanz für diese Variable bei ABBY. Im Vergleich mit Acrobat ist der Unterschied sogar stärker, was an der Cluster-Bildung im rechten Rand zu erkennen ist. ABBY verliert also bei zahlreichen Blättern keine Zeichen, bei denen Acrobat überhaupt kein Zeichen findet, also das ganze Blatt „verliert“. Der Vergleich von Omnipage und Acrobat ist ambivalenter. Vereinzelte Punkte im Quadranten I weisen auf annähernd gleich schlechte Verlustraten für diese Liedblätter hin. Eine mittelgroße Clusterbildung im Quadranten IV bestätigt jedoch wieder die Annahme der Hypothese 1.3 für Lost In Percent.

Abschließend werden noch die Scatterplots für CER gezeigt. Es gilt zu beachten, dass der CER-Wert über 100 Prozent hinausgehen kann und dies für Omnipage und Acrobat X Pro auch punktuell tut. Für die grafische Darstellung sind diese Ausreißer jedoch vernachlässigbar und der Wertebereich wurde auf 100% beschränkt:

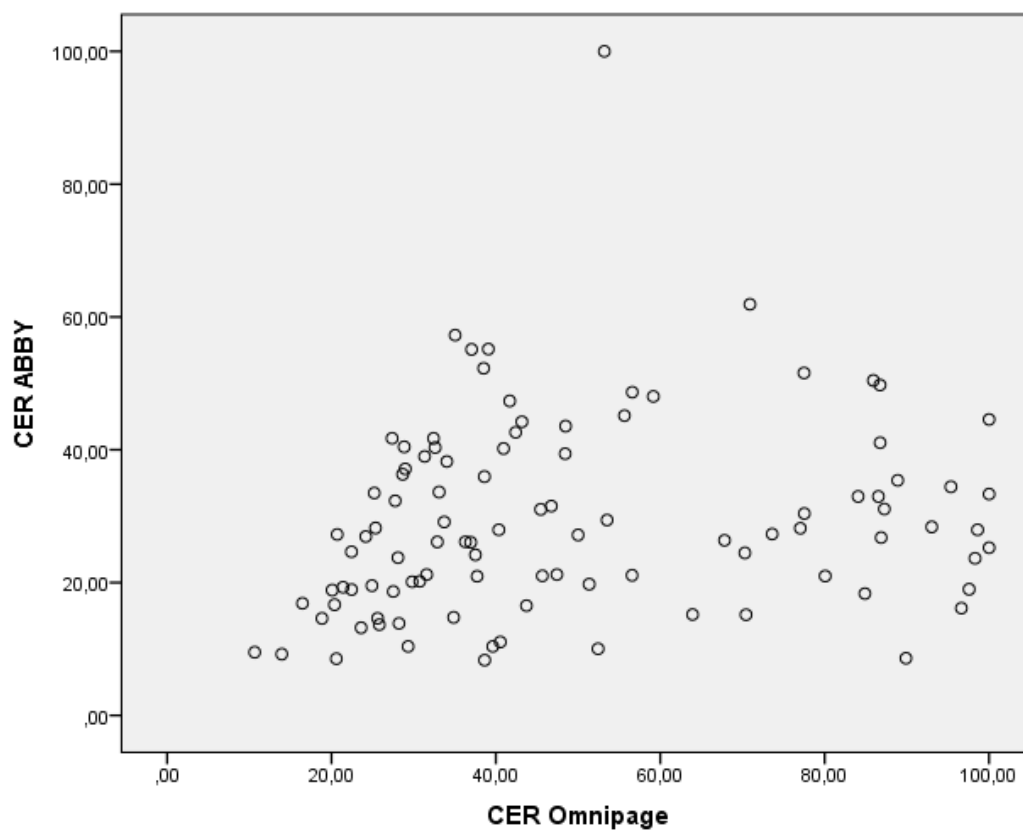


Abbildung 50: Scatterplot – CER ABBY/Omnipage

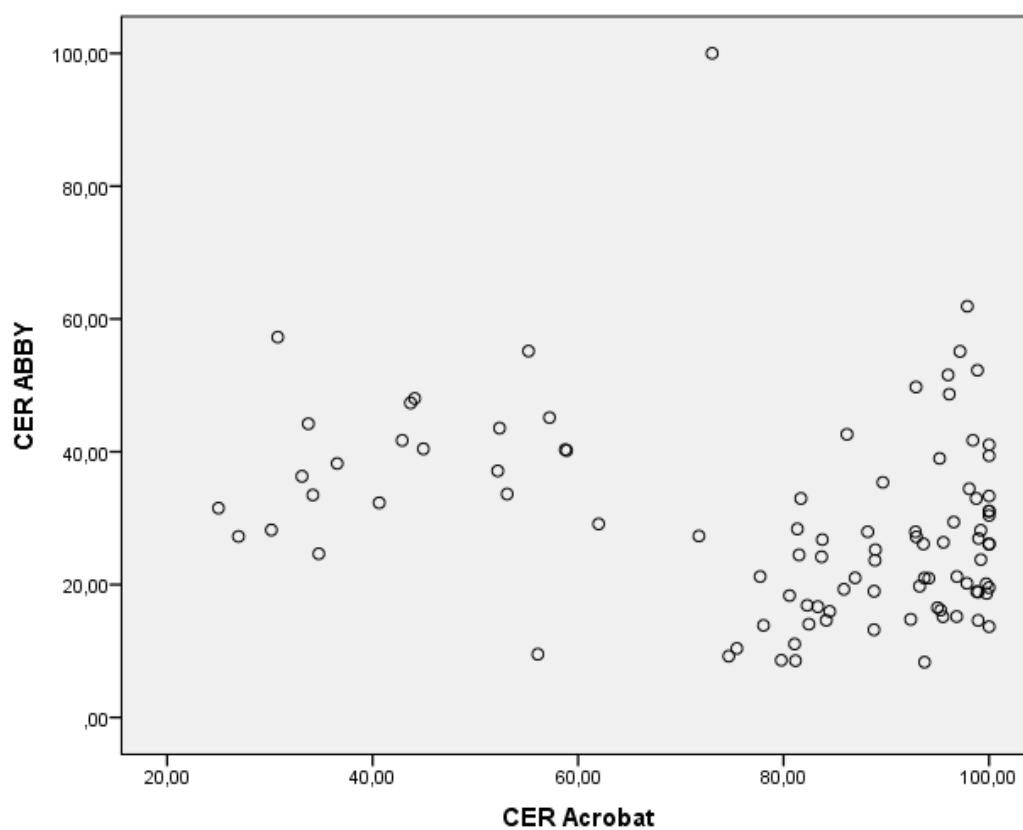


Abbildung 51: Scatterplot – CER ABBY/Acrobat

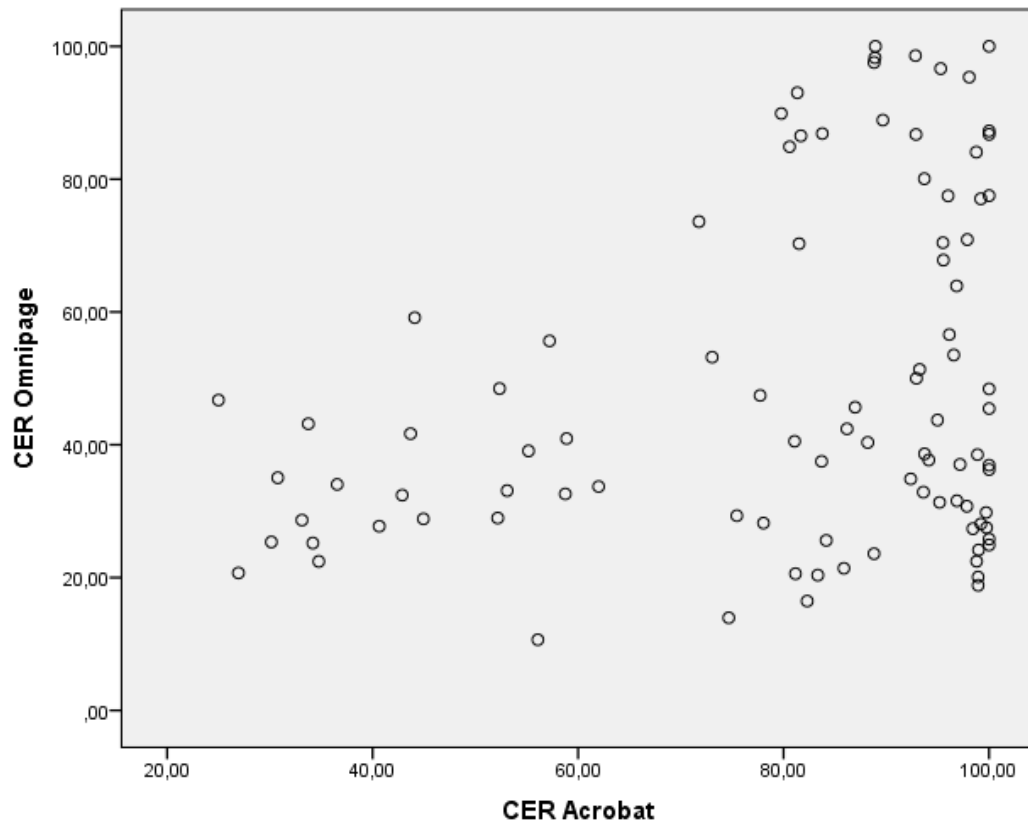


Abbildung 52: Scatterplot – CER Omnipage/Acrobat

Die angenommenen Hypothesen für die Character Error Rate können mit den Scatterplots verdeutlicht werden. Da die CER alle Fehlerarten miteinberechnet, ist es ein kritischeres Maß als die bisherigen. Das sieht man auch bei der ersten Grafik. Zwar befindet sich die Merzhzahl der Blätter im 1. Quadranten, was auf eine gute Performance hinweist. Jedoch ist die Tendenz hin zum 4. Quadranten sehr deutlich. Die bessere Leistung von ABBY ist hier also erkennbar. Im Vergleich zu Acrobat ist dies auch ersichtlich obwohl einige Werte in die Quadranten I und II streben, was punktuell „gleichschlechte“ Leistungen nahelegt. Bei Omnipage und Acrobat befinden sich mehr Blätter in diesem Bereich. Dennoch ist ein Cluster im Quadranten IV zu erkennen, was die angenommene Hypothes H1.3 auch bestätigt.

Die Scatter-Plots verdeutlichen die Hauptaussage, dass ABBY auf dem Test-Korpus eine bessere Leistung erbringt als alle anderen Tools.

3.5.11.2 Gestapelte Balkendiagramme

Folgende gestapelte Balkendiagramme visualisieren die Verteilung der Grounded-Truth-Interpretation der Tools im Vergleich. Dafür werden die absoluten summierten Werte für die totale Zeichenanzahl, die korrekten Zeichen davon, die falschen und die

verlorenen Zeichen genommen. Die Grafiken sind nur Vergleiche der in der deskriptiven Statistik der Tools bereits aufgezeigten Kreisdiagramme. In den Balkendiagrammen stehen die absoluten Zahlen der Zeichen von diesem Typ:

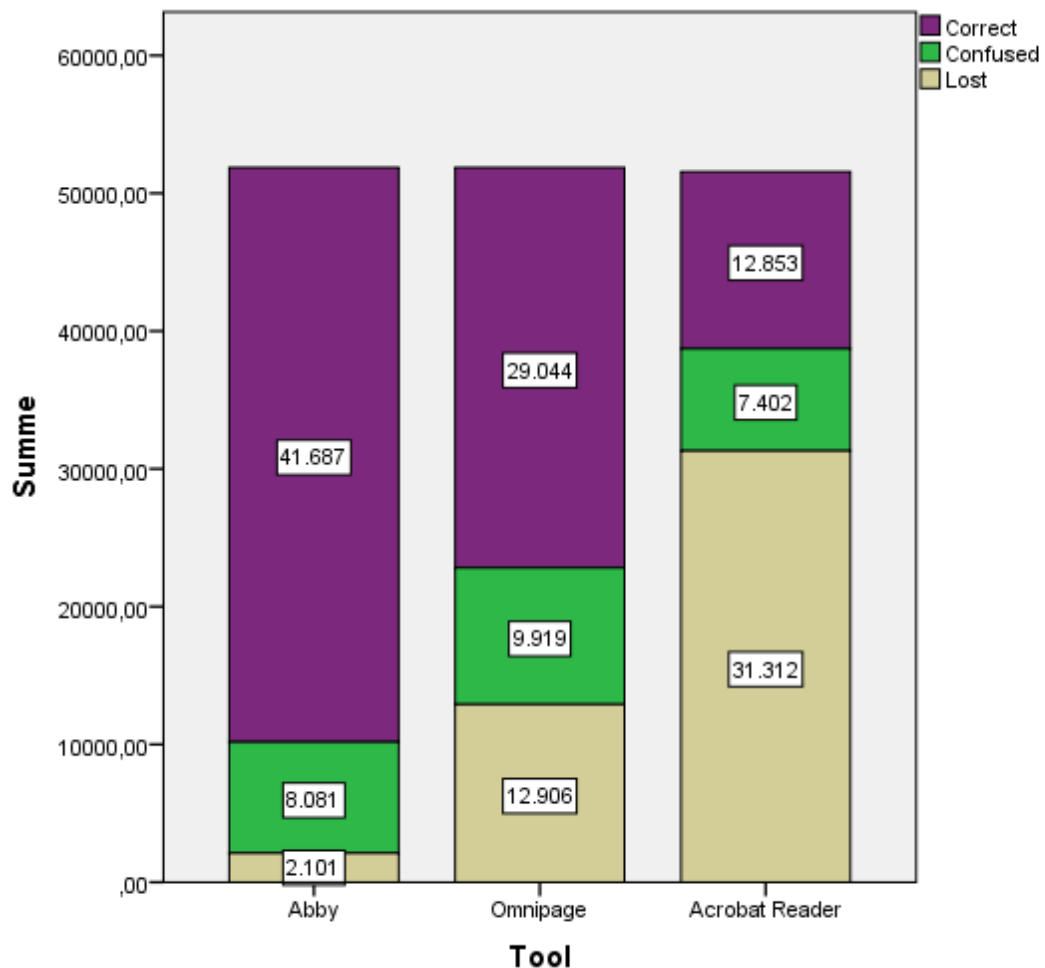


Abbildung 53: Gestapeltes Balkendiagramm Zeichentypen – Grounded Truth

Bezogen auf den Grounded-Truth-Text wird deutlich, dass ABBY deutlich mehr Zeichen korrekt erkennt. Der große Anteil an gar nicht erkannten Zeichen (>50%) weist auf die schwache Leistung von Acrobat hin. Betrachtet man dieselbe Grafik nur für den OCR-Output (so dass verlorene Zeichen nicht gemessen werden) wird deutlich wie viel kleiner der Output von Acrobat und Omnipage tatsächlich ist, weil Zeichen verloren gehen und gar nicht erkannt werden:

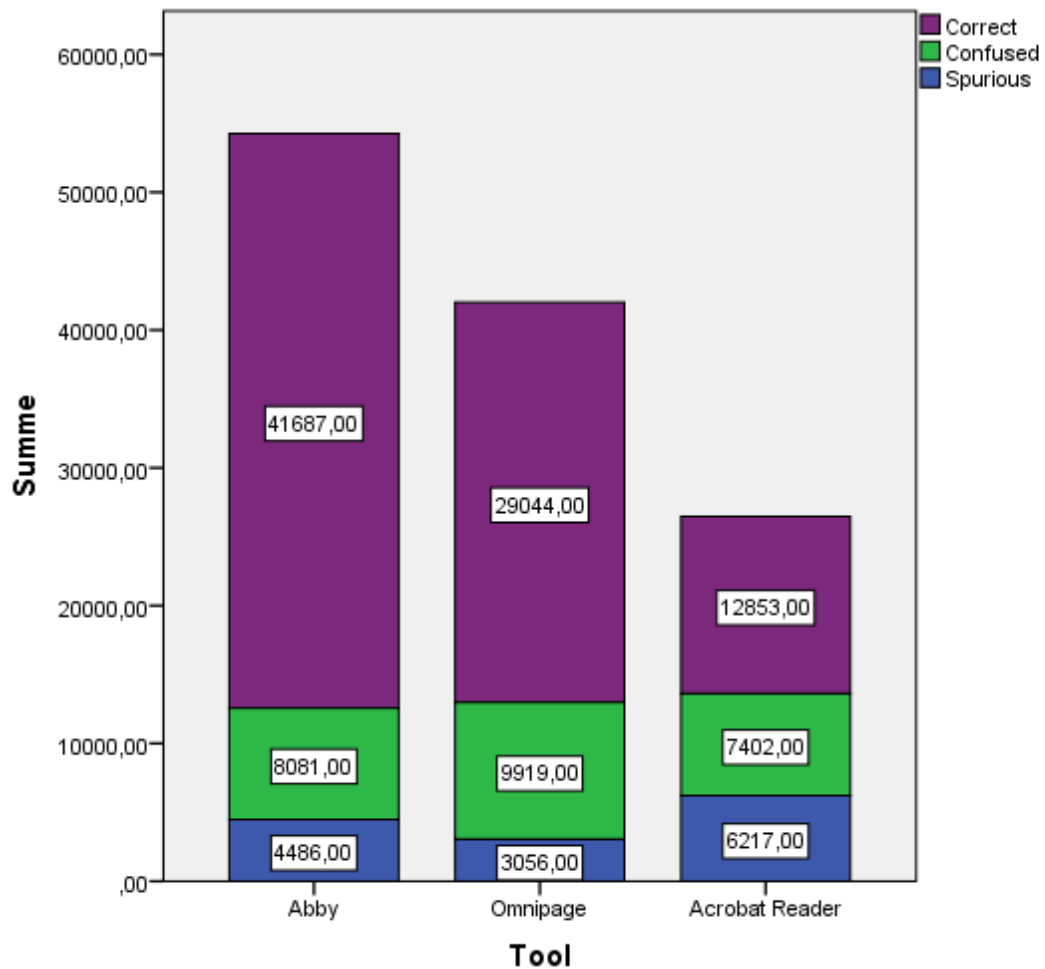


Abbildung 54: Gestapeltes Balkendiagramm Zeichentypen – OCR-Output

Betrachtet man den OCR-Output als Ganzes mit korrekten Zeichen und allen Fehlertypen wird jedoch auch ersichtlich, dass der Anteil von Spurious-Zeichen relativ gering ist. Tatsächlich hat Omnipage die geringste Noise. Die großen Verlusten bei Omnipage und Acrobat sprechen aber auch hier dafür, dass ABBY für den Testkorpus die beste Leistung erbringt:

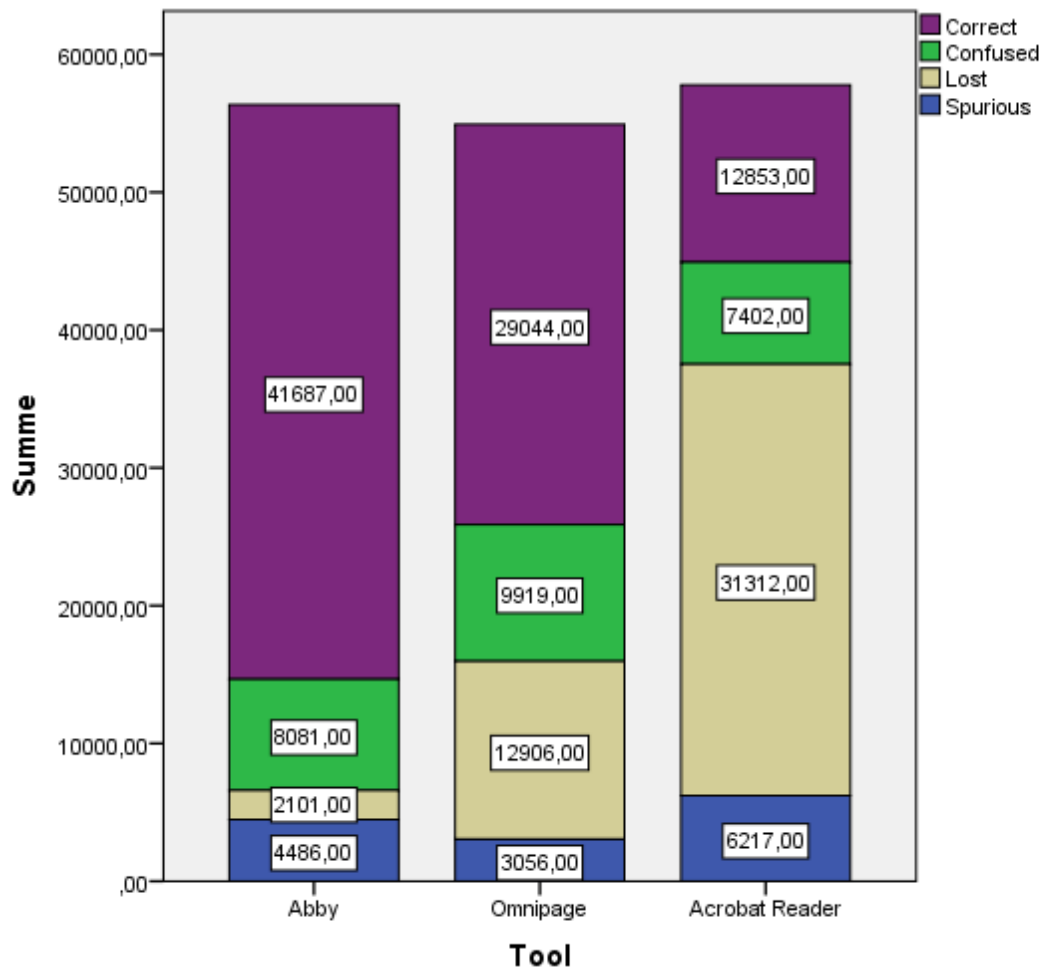


Abbildung 55: Gestapeltes Balkendiagramm Zeichentypen – Complete

4 Diskussion

Mit der vorliegenden Studie wurden zwei Ziele verfolgt. Zum einen sollte die Frage beantwortet werden, ob eine maschinelle Erfassung der textuellen Bestandteile der Hoerburger Liedblattsammlung mittels Tool-Support überhaupt Sinn macht, sowie zweitens, welches Tool dazu am besten geeignet ist.

Nach Durchführung einer vergleichenden OCR-Evaluation wurde nachgewiesen, dass ABBY in Bezug auf alle relevanten OCR-Metriken (Erkennungsrate, Precision, Character Error Rate) am besten abschneidet. Als OCR-Tool zur textuellen Erschließung kann auf Basis der vorliegenden Daten sowie der Auswertung aller getesteten Tools also ABBY empfohlen werden.

Es bleibt noch die Frage zu beantworten, inwiefern sich ein OCR-Einsatz überhaupt lohnt. Die OCR-Firmen geben meist an, dass ihre Tools eine Erkennungsrate von 99%

haben. In anderen Studien war die Performanz aller Tools deutlich besser als in der vorliegenden. Bei Boschetti, Romanello, Babeu, Bamman und Crane (2009) hat ABBY eine Erkennungsrate von 94%. Der Korpus besteht aus griechischen Büchern, die Herausforderung ist dabei die Zeichenkomplexität der griechischen Sprache. Vice, Kanai und Nartker (1992) weisen bei ihrer Studie über normale Bilddateien einer Handelsgesellschaft eine Erkennungsrate von 97% für Omnipage nach. Für Acrobat X Pro sind noch keine äquivalenten Studien bekannt. Die Ergebnisse gestalten sich jedoch deutlich unterschiedlich zu den hier erhaltenen Erkennungsraten von ca. 80% bei ABBY, 56% bei Omnipage und 26% bei Acrobat X Pro. Gemäß Holley (2009) handelt es sich in jedem Fall um „Poor Performance“. Für die anderen Metriken sind keine Empfehlungen bekannt, doch auch für die andere Hauptmetrik, die Character Error Rate, muss man das Ergebnis als schlecht interpretieren.

Holley (2009) weist darauf hin, dass die Erkennungsraten von 99% der Firmen nur bei sehr sauberen Bildern mit sehr guter Qualität erreicht werden. Bei seiner Studie über Zeitungsartikel konnte er feststellen, dass die Genauigkeit mit höherem Alter des Artikels abnimmt. Holley (2009) gibt auch an, dass ein tatsächlicher Konsens darüber, was gutes oder schlechtes OCR ist, so nicht existiert und diese Frage immer in Abhängigkeit vom Ausgangsmaterial zu betrachten ist. In einer umfangreichen Studie von Edwin (2008) für die holländische Nationalbibliothek hält er die OCR-Situation in Bezug auf historische Quellen wie die hier vorliegenden Liedblätter sehr gut fest:

“Accuracy rates, on either word or character level, should not be considered as watertight performance indicators for OCR software. Usually the quality of the OCR text says more about the condition of the original materials than it does about the performance of the OCR software. For what its worth, the rates respondents gave for newspaper digitisation projects vary from 99.8% for 700,000 newspaper pages (word accuracy, manually corrected) to 68% (character accuracy, no correction) for 350,000 pages of early 20th century newspapers.”

Der Liedblattkorpus ist schlicht nicht ideal für OCR-Erkennung. Sehr viel Noise, Verschmutzung, eine hohe Heterogenität, unterschiedliche Zeichensätze, oft sehr schlechter Kontrast, Dialekte, Fraktur-Schrift, Überschneidungen von Textkomponenten sind einige Nachteile des Korpus. Unter Betrachtung all dieser Begrenzungen kann man 80% als akzeptables Ergebnis annehmen.

Unterschiedliche Wege können beschritten werden, um das Ergebnis zu verbessern. Holley schlägt bis zu 13 mögliche Aktionen vor. Diese zeigen auch die Grenzen der hier

vorliegenden Studie auf. Es wurden keine vorhergehenden Optimierungsvorgänge über Bildbearbeitungsprozesse durchgeführt. Innerhalb eines potentiellen Workflows könnte dies jedoch den Digitalisierungsaufwand unverhältnismäßig erhöhen. Als Bildmaterial wird tiff empfohlen, das Universitätsarchiv hat die tiff-Dateien optimiert und dann als jpeg ausgegeben, was folglich den Testkorpus gebildet hat. Andere Tools, auch Open-Source-Tools, wie z.B. Tesseract könnten damit bessere Ergebnisse erzielen. Die Tools wurden ohne größere Einstellungen verwendet. In der Tat bietet aber z.B. ABBY die Möglichkeit eines Trainings an, was die Leistung verbessern könnte. Bei der Heterogenität des Datenbestands ist es jedoch zweifelhaft, wie zielführend dies ist. Weitere Möglichkeiten werden bei Holley (2009) beschrieben. Diese würden die Dimensionalität der statistischen Analyse erhöhen. Weitere Studien (möglicherweise auch in diesem Projekt) können die eben beschriebenen Aspekte genauer untersuchen.

Als eine andere Möglichkeit den OCR-Output zu verbessern beschreibt Holley (2009) die händische Auszeichnung. Nach Durchführung der OCR-Studie und nach Einschätzung des Ergebnisses wird empfohlen, diese Möglichkeit zu verfolgen. Mit 80% Erkennungsrate liefert ABBY in Bezug auf den Korpus eine ausreichende Genauigkeit nach subjektiver Einschätzung, um eine Textgrundlage für die Liedblätter zu bilden. Im Rahmen des Projekts soll jedoch ein Crowdsourcing-Tool entwickelt werden, mit dem es effektiv möglich sein soll, den OCR-Output von ABBY von den Liedblättern zu verbessern.

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.
- Bortz, J. (2005). *Statistik für Human- und Sozialwissenschaftler*. (6. Auflage). Berlin: Springer-Verlag.
- Boschetti, F., Romanello, M., Babeu, A., Bamman, D. & Crane, G. (2009). Improving OCR Accuracy for Classical Critical Editions. In Agosti, M., Borbinha, J.L., Kapidakis, S., Papatheodorou, C. & Tsakonas, G. (Hrsg.), *ECDL 2009* (S. 156-167). Berlin: Springer-Verlag.
- Carrasco, R. C. (2014). An open-source OCR evaluation tool. In: *DATeCH 2014*. New York: ACM Press.
- Fitzpatrick, J. (2010). *Five Best Text Recognition Tools*. Retrieved from <http://lifesacker.com/5624781/five-best-text-recognition-tools>
- 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>
- Kanungo, T., Marton, G. & Bulbul, O. (1999). *Performance Evaluation of Two Arabic OCR Products*. Retrieved from <http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.78.9411&rep=rep1&type=pdf>
- Karaoglu, S., van Gemert, C. J. & Gevers, T. (2012). Object Reading: Text Recognition for Object Recognition. In Fusiello, A., Murino, V. & Cucchiara, R. (Hrsg.), *Computer Vision – ECCV 2012. Workshops and Demonstrations* (S. 456-465). Berlin: Springer-Verlag.
- 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>
- Mello, C. A. B. & Lins, R. D. (2012). A Comparative Study on OCR Tools. In: *Vision Interface '99* (S. 224-232). Trois-Rivières, Canada. Retrieved from <http://www.image-ware.com.br/download/OCRE99.pdf>
- Rasch, Frieze, Hofmann & Naumann (2010). *Quantitative Methoden. Band 2. Kapitel 7: Varianzanalyse mit Messwiederholung*. Heidelberg: Springer. Retrieved from http://quantitative-methoden.de/Dateien/Auflage3/Band_II/Kapitel_7_SPSS_Ergaenzungen_A3.pdf
- Rice, S.V., Kanai, J. & Nartker, T.A. (1992). *A Report on the Accuracy of OCR Devices*. Retrieved from http://www.expervision.com/wp-content/uploads/2012/12/1992.A_Report_on_the_Accuracy_of_OCR_Devices.pdf
- Top Ten Reviews. (2015). *OCR Software Review REVIEWS AND COMPARISONS*. Retrieved from <http://ocr-software-review.toptenreviews.com/>

Anhang

Evaluation

- Grounded Truth (txt-Dateien)
- HTML-Outputs (html-Dateien für ABBY, Omnipage, Acrobat X Pro)
- OCR-Output-Bereinigt auf Textzonen (txt-Dateien für ABBY, Omnipage, Acrobat X Pro)
- OCR-Output-Rohdaten (txt-Dateien für ABBY, Omnipage, Acrobat X Pro, hier auch leere Seiten, ohne jegliche Bereinigung, alle Textzonen)
- Test-Korpus-Jpegs (alle Liedblätter als jpegs)

Programme

- ocrevalUation-1.3.0-jar-with-dependencies (Programm von Carrasco, 2014)
- OCRHtmlReader.zip (Java-Projekt des eigenen Programms)

SPSS Daten und Auswertung

- Ausgangstabellen (unterschiedliche SPSS-Tabellen für unterschiedliche Anwendungsfälle)
- Deskriptive Statistik (als SPSS-Viewer-Dateien)
- Inferenzstatistik (als SPSS-Viewer-Dateien)
- Vergleichsgrafiken (als SPSS-Viewer-Dateien)