Finde den Unterschied

Wortzahlen, Segmentzahlen
und Trefferquoten (Match-Werte)
in TM-Systemen

tekom/tcwold 2019

Angelika Zerfaß

zerfass@zaac.de

tekom Jahrestagung 2019
STUTTGART, 12. – 14. NOVEMBER

tcworld conference 2019
STUTTGART, NOVEMBER 12 – 14

Ihre Meinung ist uns wichtig! Sagen Sie uns bitte, wie Ihnen der Vortrag gefallen hat. Wir freuen uns auf Ihr Feedback unter

http://lt06.honestly.de

oder scannen Sie den QR-Code

Das Bewertungstool steht Ihnen auch noch nach der Tagung zur Verfügung!

Wann fallen Unterschiede auf?

- Zählung erfolgt mit unterschiedlichen Programmen (Word vs. TM-Programm).
- Gleiches Dokument und gleiches TM ergeben in 2 verschiedenen TM-Programmen unterschiedliche Werte.
- Gleiches Dokument und gleiches TM ergeben in 2 Versionen des gleichen TM-Programms unterschiedliche Werte.
- Gleiches Dokument, gleiches TM, gleiches TM-Programm ergibt auf verschiedenen Rechnern unterschiedliche Werte.

Unterschiede

- Woher kommen die Unterschiede bei
  - der Anzahl der Wörter,
  - der Anzahl der Segmente,
  - der Trefferquote zwischen Segmenten im Dokument und Segmenten im TM?
- Generell:
  - Unterschiedliche Zählweisen
  - Unterschiedliche Einstellungen
  - Unterschiedliche Programm-Versionen

Wörter

- Jedes Programm, das Wörter zählt, hat eine (eigene) Definition, was ein Wort darstellt.
- Bei "normalen" Texten mag kein Unterschied in der Wortzählung in verschiedenen Programmen auffallen, aber bei Texten, die spezielle Dinge beinhalten, werden plötzlich unterschiedliche Wortzahlen angezeigt.

WAS IST EIN WORT?

Was ist ein Wort?

- Zeichenfolge, die von Trennzeichen, wie z.B. Leerzeichen umgeben ist.
- Andere Zeichen werden eventuell von Programm zu Programm unterschiedlich definiert:
  - Schrägstrich /
  - Bindestrich –
  - Apostroph ', Gleichheitszeichen =, Plus-Zeichen +…

Wortzahlen

Text	Word	Programm A	Programm B
2018-07-20T10:53:26.563Z	1	3	1
image/jpeg	1	2	1
fabric-pouch-detail-4_3-1240.jpg	1	5	1
Satz mit « französischen Anführungszeichen » im Text.	8	6	6

Programm A:
5 Wörter – aber mit welchem Trennzeichen?

fabric-pouch-detail-4_3-1240.jpg

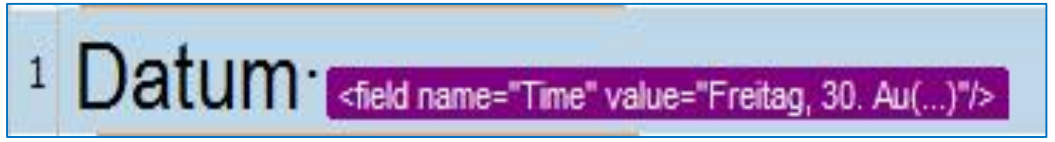
fabric-pouch-detail-4_3-1240.jpg

- Bindestriche sind nur Trennzeichen, wenn Zahlen im Spiel sind.
- Punkte sind nicht immer Trennzeichen.
  - 123.jpg = 1 Wort
  - 123_123 = 2 Wörter
  - 123_123.jpg = 2 Wörter (Unterstrich ist Trenner)
  - 123-123.jpg = 3 Wörter (Bindestrich und Punkt sind Trenner)

Wortzahlen

Segment	Word	Programm A	Programm B
Satz mit einer Referenz (siehe Seite 1).	7	6, wenn Zahl ein Feld ist 7, wenn Zahl Text ist	7, egal ob die Zahl Text oder Feld ist
Satz mit einem Link.	4	4, wenn Text nur blau unterstrichen ist 5, wenn z.B. eine URL hinterlegt ist	4, wenn Text nur blau unterstrichen ist 5, wenn z.B. eine URL hinterlegt ist

- Die Art und Weise, wie Text geschrieben wird, beeinflusst die Wortzahl.



- Darstellung von Feldinhalten unterscheidet sich.

Wortzahlen

- Unterschiede innerhalb eines Programms, je nach Einstellungen.

Einstellung im TM	Texte	Wortzahl
Texte analysiert ohne TM bzw. mit Standardeinstellungen im TM	Maßeinheit ohne Leerzeichen – 25km	4
	Maßeinheit mit Leerzeichen – 25 km	4
Texte analysiert mit TM, Maßeinheitenerkennung aus	Maßeinheit ohne Leerzeichen – 25km	4
	Maßeinheit mit Leerzeichen – 25 km	5

Wortzählung

- Die Anzahl der Wörter in derselben Datei kann in verschiedenen (TM-)Programmen unterschiedlich sein.
- Diese Einstellungen bestimmen, wie viele Wörter das TM-Programm "sieht":
  - Versteckte Texte/Ebenen/Folien
  - Geschützte Texte/Tabellenblätter
  - Kommentare, Notizseiten (PPT)
  - Eingebettete Dateien
  - Texte zwischen oder innerhalb von Tags (Attributwerte)
  - Spezielle Einstellungen, die Tags als Wörter zählen können, um den Übersetzer auch für die Arbeit mit den Tags zu entlohnen.
- Bei Zählungen während eines Projekts können gesperrte Segmente gezählt oder ignoriert werden.

Wortzählung

- Dateiformate können auch Einfluss auf die Wortzählung haben. Zum Beispiel: Zählung einer PDF-Datei und der Originaldatei aus InDesign, Word…
  - Die Wortzahl der PDF-Datei ist höher, weil hier alle sichtbaren Texte gezählt werden (Inhaltsverzeichnis, Index, jede Kopf- bzw. Fußzeile wird separate gezählt).
  - In der Originaldatei werden nur die Überschriften gezählt, die später das Inhaltsverzeichnis ergeben oder der Text einer Kopfzeile wird nur einmal gezählt, obwohl er auf mehreren Seiten sichtbar ist.

SEGMENTE

Segmente

- Ein Segment ist Teil eines Dokuments.
- Segmente werden anhand von Segmentierungsregeln erkannt.
- Ein Segment kann sein:
  - Einzelne Zeichenfolge (z.B. eine Überschrift die nur aus 1 oder 2 Wörtern besteht)
  - Teilsatz (z.B. Punkte einer Aufzählungsliste)
  - Satz
  - Absatz (wenn die Segmentierung nur an der Absatzmarke vorgenommen wird)

SEGMENTIERUNGSREGELN

Segmente

- Punktregel
  - Ein Punkt ist ein Segmentende, wenn danach ein Leerzeichen folgt und danach ein Großbuchstabe oder eine Zahl.
- Ausnahmen zur Punktregel
  - Wenn der Punkt einer bekannten Abkürzung folgt, ist der Punkt kein Segmentende (z.B.: Dr. Müller).
- Ähnliche Regeln für andere Satzendezeichen (!, ?, :)

Segmente

- Regeln für Begrenzungen…
  - Harte Zeilenschaltung = Segmentende
  - Zellenbegrenzung / Rahmen einer Textbox = Segmentende
  - Weiche Zeilenschaltung/Tabulator = normalerweise kein Segmentende (kann aber oft benutzerspezifisch im TM-Programm eingestellt werden).

Segmente

- Regeln bei getaggten Dateien
  - Ein Struktur-Tag ist ein Segmentende.
  - Ein Inline-Tag ist kein Segmentende.
  - Text zwischen Tags (un)übersetzbar?
  - Text innerhalb von Attributen übersetzbar?

```
−<sample>
  <title>Erstellen eines XML Filters</title>
 −<text>
    Liste der
    <emphasis>Tag-Namen</emphasis>
    auslesen.
  </text>
  <instruction>4f</instruction>
 −<text>
    Definition der
    <emphasis>Struktur-Tags</emphasis>
    und
    <emphasis>Inline-Tags</emphasis>
    vornehmen.
  </text>
  <instruction>5r</instruction>
 −<text>
    Prüfen, ob es Text zwischen Tag-Paaren gibt, der nicht übersetzt werden darf.
  </text>
  <instruction>6m</instruction>
 −<text>
    Prüfen, ob es Text innerhalb von Tags gibt (Attributwert), der übersetzt werden muss.
  </text>
  <button value="Abbrechen"/>
  <button value="OK"/>
</sample>
```

Segmente

- Anzahl der Segmente richtet sich nach:
  - Elementen im Ausgangstext und deren Definition im TM-Programm
    - Harte Zeilenschaltungen an Stellen, an denen sie Sätze trennen (z.B. in Textboxen oder Tabellenzellen)
    - Tabulator und weiche Zeilenschaltung (Segmentende oder kein Segmentende)
  - Einstellungen zum Import von Dokumentinhalten
    - Versteckte Texte, Ebenen
    - Eingebettete Inhalte
    - Kommentare, Notizseiten, Dateiinformationen (Metadaten)
    - Inhalte von Variablen

TREFFERQUOTEN (MATCHES)

Trefferquoten

- Die Trefferquote im TM-Programm bezeichnet die Ähnlichkeit eines Segments im Dokument mit einem Segment im TM (in Prozent).
- Jedes Programm hat seine eigene Art und Weise, wie diese Prozentzahl errechnet wird.
  - -> Das gleiche Dokument mit dem gleichen TM in unterschiedlichen TM-Programmen zeigt unterschiedliche Trefferquoten an.

Trefferquoten

- Vergleich des ersten Segments mit dem zweiten und des zweiten mit dem dritten in 2 verschiedenen TM-Programmen.

1.	Dies ist ein neuer Satz.	This is a new sentence.	0%
2.	Dies ist ein kurzer neuer Satz.	This is a short new sentence.	73%
3.	Dies ist ein kurzer schöner Satz.	This is a short new sentence.	70%

	Demo 1.docx		Demo 1.docx
1	Dies ist ein neuer Satz.		This is a new sentence.
2	Dies ist ein kurzer neuer Satz.	80%	This is a short new sentence.
3	Dies ist ein kurzer schöner Satz.	89%	This is a short new sentence.

Trefferquoten

- Programm 1 reserviert Trefferquoten zwischen 95% und 99% für Segmente, die sich nicht im Text sondern nur in anderen Dingen unterscheiden (z.B. Formatierung, Tags, Zahlen, Groß-/Kleinschreibung, Leerzeichen)
- Programm 2 würde bei Textänderungen auch Trefferquoten zwischen 95% und 99% anzeigen

99% - Tag

99% - Zahl

98% - Tag + Zahl

98% - Zahl und Formatierung

Einfluss auf die Trefferquoten

- Die Einstellungen zur Segmentierung sollten für aufeinanderfolgende Übersetzungen dieselben sein, um bessere Trefferquoten zu erreichen.
- Kleine Änderungen zwischen Dokumentversionen, z.B. das Auftrennen eines Satzes in zwei einzelne Sätze, verringern die Trefferquoten und erhöhen die Kosten.
- Unbekannte Abkürzungen und unnötige Zeilenschaltungen führen im TM-Programm falscher Segmentierung, die der Übersetzer manuell ändern muss.
- Unterschiedliche Einstellungen im gleichen TM-Programm auf verschiedenen Rechnern (Abzüge)
- Umstieg von einem TM-Programm auf ein anderes.
- Umstellung von einem Dateiformat auf ein anderes, z.B. von Word auf XML.

Unterschiede im gleichen TM-Programm

- Abzüge auf Treffer im TM können die Trefferquoten beeinflussen.
  - Abzug auf Treffer aus einem Alignment
  - Abzug auf Treffer mit bestimmten Metadaten
  - Abzug auf Treffer aus einem bestimmten TM

Name	Aktiviert	Sprachen	Suche	Abzug	Konkordanz	Aktualisieren
Kunde A_allgemein.sdltm	☑	DE → US	☑	0	☑	☐
Kunde A_Marketing.sdltm	☑	DE → US	☑	5	☑	☐
Kunde A_Verträge.sdltm	☑	DE → US	☑	0	☑	☑

Unterschiede im gleichen TM-Programm

- Einstellungen zum Erkennen von ähnlichen Segmente
  - Zwischen Dokument und TM (Standard)
  - Zwischen Segmenten im Dokument (Zusatzeinstellung)

Berechnung der Ähnlichkeiten innerhalb eines Dokuments

Programm A:
Satz 1 und Satz 2 = ein neuer Satz und eine Wiederholung.

Text

Dies ist Satz 1.

Dies ist Satz 2.

Dies ist ein neuer Satz.

Das ist ein neuer Satz.

Dies ist auch ein neuer Satz.

Dies ist der letzte Satz im Dokument.

6 neue Sätze (bei Standardanalyse)

Programm A

Gesamtüberblick

Gesamt	Typ	Segmente	Wörter
Dateien:1	PerfectMatch	0	0
Zeichen/Wort:3.87	Kontext-Match	0	0
	Wiederholungen	1	4
	100 %	0	0
	95% - 99%	0	0
	85% - 94%	0	0
	75% - 84%	0	0
	50% - 74%	0	0
	Intern:		
	95% - 99%	0	0
	85% - 94%	0	0
	75% - 84%	2	11
	50% - 74%	0	0
	Neu/MÜ	3	16
	AdaptiveMT-Baseline	0	0
	AdaptiveMT mit Lerneffekten	0	0
	Gesamt	6	31

Programm B

Analyse

Bereich *Ausgewählte Dokumente, Anzahl der Dokumente: 1*
Ressourcen *Homogenität*

Typ	Segmente	Ausgangswörter
Alle	6	31
Dokumentenbasiert vorübersetzt / doppelter Kontext	0	0
Wiederholung	0	0
101%	0	0
100%	0	0
95%-99%	1	4
85%-94%	0	0
75%-84%	1	5
50%-74%	1	6
Kein Treffer	3	16

Programm B: Satz 1 und Satz 2 = ein neuer Satz und ein 99% Match.

Umstieg von einem TM-Programm aufs andere

- Zusätzlich zur unterschiedliche Berechnung der Trefferquoten gibt es noch weitere Unterschiede.
  - Verschiedene Art und Weise Kontext zu speichern (Beim Austausch über TMX werden Kontext-Treffer zu (maximal) 100%-Treffern.

```
<prop type="x-Context">7804750128609610620, -32094929311248036 78</prop>
<prop type="x-ContextContent">Segment eins. |  | Segment one. | </prop>
<tuv xml:lang="de-DE">
<seg>Segment zwei.</seg>
</tuv>
<tuv xml:lang="en-US">
<seg>Segment two.</seg>
</tuv>
```

Programm A
Codes + Segment und Übersetzung davor

```
<tuv xml:lang="de-de">
<prop type="x-context-pre">Segment eins.</prop>
<prop type="x-context-post">Segment drei.</prop>
<seg>Segment zwei.</seg>
</tuv>
<tuv xml:lang="en-us">
<seg>Segment two.</seg>
</tuv>
```

Programm B
Ausgangssegment davor und danach

Die Frage nach einem Standard

- Die Berechnung von Wortzahlen und Trefferquoten sind grundlegende Merkmale eines TM-Systems. Eine Standardisierung über TM-Systeme hinweg ist nicht gewünscht.

- GMX-V (https://www.gala-global.org/gmx-v-10)
- GMX-V addresses the issue of quantifying the workload for a given localization or translation task. This is often commonly referred to as word counts. Word counts, however, do not convey the true range of possible metrics that can be used to assess the cost of localizing a document such as the number of screen shots for a software localization project, or page counts for a document layout task. GMX-V is a more precise definition of the metrics required for billing and sizing purposes.

tekom Jahrestagung 2019
STUTTGART, 12. – 14. NOVEMBER

tcworld conference 2019
STUTTGART, NOVEMBER 12 – 14

Ihre Meinung ist uns wichtig! Sagen Sie uns bitte, wie Ihnen der Vortrag gefallen hat.
Wir freuen uns auf Ihr Feedback unter

http://lt06.honestly.de

oder scannen Sie den QR-Code

Angelika Zerfaß
zerfass@zaac.de

Das Bewertungstool steht Ihnen auch noch nach der Tagung zur Verfügung!