How were paraphrases, allusions or translations reused by authors over time?

Why was part of the text reused? What influenced the text?

Aim
Develop a methodology to „measure“ historical text reuse in spite of its diversity

Through
Big Data, texts in Ancient Greek, German, English, Italian, Latin
Information overload (mass digitization) vs. information poverty (many lost works). Incomplete cut-out of what actually exists & existed

Visualization Example
Bible graph visualization (TRAViz) to show variants of verses.

Image by Stefan Jänicke
ACIDiversity

- **Stability (yellow):** syntactic vs. semantic
- **Purpose/Intention (green):**
- **Size of Text Reuse (blue):**
- **Literary classification (light blue):**
- **Degree of distribution (purple):**

**Challenge**

Language Modeling to cope with this diversity
Direction detection (not just symmetric relation)
Boundary detection (not just point to the sentence)

Challenge

Realistic time management. Be aware of the complexity of a task.
Different **formats**: annotation; encoding.

**Challenge**
How to **homogenize** datasets in order to ensure interoperability and **minimize** data preparation?

<table>
<thead>
<tr>
<th>cit-quote-bibl</th>
<th>blockquote</th>
<th>bibl without quote</th>
</tr>
</thead>
<tbody>
<tr>
<td><code>&lt;cit&gt;</code></td>
<td><code>&lt;quote rend=&quot;blockquote&quot;&gt;</code></td>
<td><code>&lt;p&gt;</code></td>
</tr>
<tr>
<td><code>&amp;lt;quote&gt;</code></td>
<td><code>&lt;line&gt;</code></td>
<td>`[...]a)nti\ tou= proe/pinon. kuri/ws ga\ r e)sti tou=to propi/nein, to\ e(te/rw) pro\ e(autou= dou=nai piei=n. kai ( *)odussus\ s de\ para\ tw=</td>
</tr>
<tr>
<td><code>du/o ku/nes a)rgoi\ ei(/ponto</code></td>
<td><code>a)gxou= d' i(stame/nh e)/pea ptero/enta proshni/da</code></td>
<td><code>&lt;/quote&gt;</code></td>
</tr>
<tr>
<td><code>&lt;/bibl&gt;</code></td>
<td><code>&lt;/bibl&gt;</code></td>
<td><code>&lt;/p&gt;</code></td>
</tr>
<tr>
<td><code>&lt;/cit&gt;</code></td>
<td><code>&lt;/line&gt;</code></td>
<td><code>&lt;p&gt;</code></td>
</tr>
<tr>
<td><code>Od. 2.11</code></td>
<td><code>a)ll a)/ge nu=n ma/stiga kai\ h(ni/a sigalo/enta</code></td>
<td><code>&lt;bibl n=&quot;Hom. Il. 5.226&quot;&gt;Il. 5.226&lt;/bibl&gt;</code></td>
</tr>
<tr>
<td><code>&lt;/cit&gt;</code></td>
<td><code>&lt;/line&gt;</code></td>
<td><code>[...]</code></td>
</tr>
<tr>
<td><code>&lt;/cit&gt;</code></td>
<td><code>&lt;/quote&gt;</code></td>
<td><code>&lt;/quote&gt;</code></td>
</tr>
</tbody>
</table>
‣ Which are the determinable **minutiae** of text reuse and which of them can be **calculated** by **machines**?

‣ **What can’t** be calculated?

‣ How can text reuse analysis be improved by the calculation of **paradigmatic relations**?
Greta Franzini, Humanist

Can the study of text reuse...

- ...tell us something about the **accuracy** and **confidence** of **variation**? Which authors quote the source more literally and why?

- ... tell us something about the **nature** of **transmission contamination**?
When observing a sentence or word that has been reused, one may ask:

- What are the **differences** and **similarities**, what do **they tell** us about the texts?
- **How** does one **classify** the similarities and the differences (**Wittgenstein**)?
- How and can those findings help to improve **NLP**?
The reuse of text across languages is a gold mine for transcultural studies. When observing translated text,

‣ ...what is the scale of divergence from one text to the other?

‣ ...what is the divergence caused by? Translator competence or linguistic, cultural, ideological and political norms?

‣ Could machine translators be used to rid a translation of any contextual influences?
Thank you for your attention!

etrap.gcdh.de