In search of formal identity and difference: considerations based on collating some Turkic manuscripts

by M. van Damme & H. Braam

1.1

Establishing differences and identities between two or more copies of the same text, generally called collating, has largely been done manually. When collating by hand one is faced with some inherent restrictions, such as the personal memory which considerably restricts the number of copies that can be taken at the same time into account. A manually performed collation has a major defect, namely that it cannot be used anew easily, nor can it be expanded after publication in print. For instance if a new copy shows up the entire collation must be redone.

The use of computers in philology is generally accepted today, and several computer procedures for collation are already available. It is our contention, however, that the revolutionary aspects of computerized philology (only to be compared with the impact of the printing press on the humanities in the Renaissance¹), are not yet fully perceived. We do not want to consider particular computer programs here, but we will try to introduce a formal way of thinking about collating, such that one can describe the manuscript material without reference to a specific program. The collating program we envisage here will provide us with a lasting apparatus criticus, independent of the text finally printed or otherwise made public. We introduce formal concepts that will make it possible to describe the material without taken a specific stance with regard to the question of the best text. This formal approach will make it possible to design thoroughly philological projects, without creating chaotic situations (representation conflicts, structural incompatibilities etc.).

1.2

What seems trivial to human intellect is far from trivial to a programmed machine. When we want to compare two copies of a text we are immediately faced with a basic problem: what is the identity that holds between text A and text B? Since we want to use a computer, we must express this notion of identity and its counterpart, difference, in such a way that they can be handled by a programmed machine. This means that we must express them formally.

In order to define formal identity and formal difference, we must always make the context explicit. The notion of context as used here must not be confused with the intuitive human notion, here we mean the notion of context as defined by criteria, chosen more or less conventionally by the users. For instance on the word level the formal identity of a specific word yoq consists in the conjunction of identities on the letter level: y, o, q in this order. But already here this simple identity does not suit our purposes, since alternative orthographies are in a sense identical, although in another sense different. It all depends on what we install as criteria. If we consider vowel length not as a criterion then yoq and 'yoq' must be considered as formal identical entities, the same counts for alternative orthographies like yoq and yoh. One could sharpen even more the notion of identity of words, if one takes into account the alternations in the diacritics of the text.

Figure 1

Here in the former form, the diacritics of the letter ِ is missing, which could have resulted in misreadings by subsequent copyists.

If we leave the word or letter level and look at the same word yoq on a higher textual level — for instance the phrase level - we are no longer comparing individual words (types), but phrases as such. Relative locations of word tokens can be defined using these phrase identities and differences. On the phrase level phrase identity could be defined using the sequences of the words and the previously defined formal identity of those words.

Example:

(1) طلقتةم بوق طلاقتةم بوق

(2) طلاقتةم بوق طلقتةم بوق

Figure 2
If we consider the vowel length irrelevant, then (1) and (2) are formal identical phrases. Should we take into account the vowel length then the phrases are said to be formal different.

The next two phrases:

\[
\text{بوق طاقميم بوق}
\]

are obviously formally different, although it is quite possible that on a higher, more interpretative level they are said to be identical since meaning the same.

Levels above words can or cannot be explicitly given by the text. If not, the collator has to fulfill an extra task: discerning and specifying phrases, sentences, paragraphs, chapters, stories etc. Sometimes programs can be used to help us, but since the segmentations are immediately coupled with the syntactical structure and the meaning of the text, much work will have to be done by the collator.

1.3

The text taken as an abstract entity is thought to be concretised in different versions, written down in a number of different MSS. Although all MSS represent the same paradigmatical text, quite a number of formal differences show up. We introduce here the notion of the texttree to describe in full the different MSS in such a way that their differences and identities can be formalised. The texttree must be considered as the most basic representation on which all other text descriptions and text manipulations depend. The notion should not be confused with the concept of a stemma. The texttree only lines up all the contents of the known MSS of a text. The hierarchy of the tree corresponds to the hierarchy of the texts of the MSS: MSS, BOOKS, LEAVES, LINES, WORDS..., not to the genetic stemmatological hierarchy. Genetic aspects based on internal affiliations are not our concern here (Cf. figure 4).

The texttree is used to be able to unambiguously point to every single word of the MSS. This is done by specifying the path to it in the tree, starting at the top (T).

Example: the second word on the 17th line on the 14th folio recto of the first MS of the collection T, can be written formally as

\[T:1:14:1:17:2.\]

In principle every single word can be pinpointed by using this texttree. A number of pinpointed words in a row can be taken as to form another, more complex entity. The character of these new, complex entities depends on the criteria that are used to divide the texts. In a lot of cases the text itself already suggests a pattern — like chapters or paragraphs — that cuts up the text into entities larger than a single word. Sometimes the reading marks (periods, colons etc.) suggest a pattern (of sentences). This cutting pattern is called a net, thrown over the basic texttree. The places where the net meets the tree define exactly the starting and ending positions of the more complex entities. A net can cut up a text in portions on all different levels of the texttree — we can impose a net on the level of lines (collation of poetry), on the level of leaves (collating convolutes) or even on the level of complete MSS (when working out a stemma). All these different nets are synchronized and interconnected by the basic texttree.

Different criteria for cutting up the texts generally imply different nets. It is part of our methodology that every formalisable systematical interpretation of textual material brings about its specific net structure. One can impose linguistic, semiological or paleographical networks, literary networks, ideological networks in so far as they can be formalized. Here a multidisciplinary treatment of the text is envisaged. For instance when collating we encounter

— on the level of MSS a network of affiliations (a stemma),
— on the level of folia a network of missing and extra folia and a network of different copyists,
— on the word level we find networks of paleographical nature and networks of legitimations, genealogical references, of poetical, ideological, narratological and linguistic nature.

2.

In order to illustrate our general remarks let us take a more concrete case. Among all possible networks of the Rabghuzi MSS we choose the narrative one.

Our text consists of series of cycli about the pre-Islamic prophets (like Adam, Salomon, Moses and Jesus). The cycli in their turn are built up from one or more autonomous stories. Each story can be divided into substories, substories into passages and passages into subpassages. The cutting up of the text into cycli, stories etc. is to a large degree given by the material texts themselves: changes of ink, ductus, titles etc. The
narrative net that corresponds to this pattern is obviously of an hierarchical nature.

As our example we choose a substory out of a story from the Noah cycle. In this substory it is told how men, women and animals entered the Ark. We compare five MSS. This substory, called here 'Entering', in its turn can be divided into four passages, as can be gathered from figure 5.

Figure 5

Narrative Net

level

- Qisas al-Anbiyaa
- of Rabghuzi
- cycles
- stories:
- substories:
- passages:
- subpassages:

Abbreviations used: Entering = The Ark's boarding procedure
Anim = The boarding of the animals
Descrip.Ark = Description of Noah's Ark
GodsOrder = God's Order to fill the Ark
How? = How to execute God's Order
Execution = The execution of God's Order
InArk = The animals in Noah's Ark
People = People learn to enter the Ark

Len5 seems to belong to a quite different redactional tradition. Len2 gives the most elaborate version. The description of how the animals entered the Ark is missing, but instead it is the only MS that has In.Ark

3.1

Every network is essentially a paradigm imposed on a text. For instance if we divide a story into substories, we impose a specific substory paradigm to the text (of the story). If we compare the texts we investigate how this paradigm is realised in other texts on the syntagmatic level. This paradigm/syntagma structure is applied on every level we are collating. For instance the narrative paradigm with respect to the passage under consideration is realised in MS Len1 as the sequence of subpassages GodsOrder, How?, Execution and in MS Len4 as How?, Execution. The syntagmatic realisation of the narrative paradigm Noah: Entering: Anim has been increased with an extra element in Len2 (Descrip.Ark).

How do we proceed here? First we must record whether there are other syntagmatic realisations of our paradigm other than those which the reference text contains. If there are, then the question arises how these realisations differ (formally with respect to the paradigm). If the realisations are completely parallel no further problems with respect to collating occur. If there is no complete parallelism then our collating activity will consist in establishing the formal differences between the syntagmatic realisations. Only when two realisations are parallel on a certain level, we can probe further, and try to establish differences and identities on a lower level. Here we meet a methodological circle: in order to collate on a lower level, we need identities on a higher level. Higher level identities however are only to be trusted when based on lower level collations. This is the reason why we can only depend on computer programs for help, whereas the formulation of hypothetical identities, with respect to hypothetical paradigms, must be done by the experienced philologist.

Formal difference and identity can be described using a limited number of basic categories, which in turn can be further subdivided. We distinguish identity, omission, addition and change.
Especially in the case of A/B some ambiguity arises which must be noted by collating programs: A/B can also be written as: A/0 followed by 0/B. E.g. an omission followed by an addition. Another ambiguity occurs with respect to the notions of addition and omission. It is completely dependent on the (conventional) choice of the reference text whether one writes 0/A (addition) or A/0 (omission). It depends on the paradigm whether one can interpret a pair as being actually an A/A pair, differing only internally or as an A/B-pair. Only when we use computer facilities these ambiguities can be kept afloat, because in the end one is not limited, nor committed to a description of the textual variants in relation to only one reference text.

Since in many cases the selection of the best MS can only be done after collation, the instance of the reference text is at best a hypothetical one. Expert systems should assist the philologist in making the best choice, possibly at any moment of the collation process.

3.2

The process of collating MSS takes into account all aspects pertaining to the MSS taken as different sequences of graphical markings, such as variants in orthography, damages, metatheses, omissions, repetitions etc. The different MSS can be seen as different realisations of the general manuscript paradigm: the capital T in figure 4. The texttree represents all realisations of a paradigm text. It is impossible to collate MSS purely on the material level, since parallel passages generally do not occur on corresponding locations in the graphical sequence of the MS texts. An addition, for instance, results in a longitudinal shift. The narrative net mentioned above offers us a formalism to express parallelism of two or more passages or subpassages. The division of a passage into subpassages is normally the result of the personal interpretation of the collator, and this interpretation may be proven to be wrong, and subsequently corrected. Their function to reveal parallel textual material is not changed by this. Once the parallel subpassages are marked, we cut them up according to a syntactically oriented net: subpassages consist of sentences, sentences consist of phrases and phrases consist of words. The next step is to collate on the sentence level, followed by the collation of phrases of corresponding sentences, and finally by the collation of words of the corresponding phrases. Collation on the graphical level belongs of course also to the possibilities. For instance in the case of Arabic scripts, one can distinguish between graphememes (completed letters) and graphemes (consonant and diacritics).

On all levels we apply the four basic collation categories: addition, omission, change and identity. For instance, a sentence can be absent from a certain subpassage, two phrases can occur in reversed order and in another phrase an extra wordform can be realised. Two words from corresponding phrases can differ in orthography etc. (See figure 2)

Figure 7. A few examples of syntagmatic relations:

- 1) A/A identity
- 2) A/0 omission
- 3) 0/A addition
- 4) A/B change

On the narrative level (subpassages) compare figure 6.
It will be clear by now that we see collation of MSS as the process of relating MSS and parts of MSS to each other. We have distinguished four basic categories of such relations. Now, how can we express such relations? Here two aspects come into account.

A) How do we describe the entities that are to be related (the fundamentals)? and B) How do we describe the relation (the connection) itself? Both aspects must be implemented in such a way that a program can handle them. We require a formally consistent and rigorous description mechanism.

The paradigmatic names used in the texttree and the nets are used to describe the specific syntagmatical entities like complete MS texts, stories, subpassages, words etc. The texttree plays a crucial role here, since it provides a unique description of all defined entities. For instance, the second word on the 17th line of the 14th folio recto of the first MS of the collection T can be described thus:

\[ T':1:14:1:17:2. \]

T:1 describes the complete MS 1 out of the collection T, T:2:15 describes a complete leaf (the 15th) of MS2 of T.

A row of entities — and most syntagmatical realisations are in such a form — is described by specifying the beginning and the end of the entity.

For instance: the passage in figure 9, the ‘Entering’ fragment of Noah in Len1 in a coded version, is described as:

\[ < T:1:14:1:9; T:1:14:2:13 >. \]

Or after deleting superfluous symbols:

\[ < T:1:14:1:9; :2:13 >. \]

In fact, this formalism is used to describe the cutting points of the nets previously discussed. This means that one could also write, describing with respect to the narrative network:

\[ N:1:Qisas:Noahcycle:Noah:Entering. \]

The subpassage ‘Execution’ (Cf. figure 5) can be written as:

\[ N:1:Qisas:Noahcycle:Noach:Entering:Anim:Execution or alternatively \]

\[ < T:1:14:1:13:7; T::14:7 >. \]

In this way all entities are described.

We want to establish between two fundamentals a relation. This is done by using a structure of the form ‘T1 R T2’. Here ‘T1’ and ‘T2’ stand for the description of the fundamental entities, and ‘R’ stands for the specification of the relation. We describe the category of a relation with one of the symbols =, +, -, *, (identity, addition, omission and change). The relations can be further qualified by adding special qualifiers. These qualifiers are dependent on the field of application. The qualifiers are added to the category symbol. For example the fact that the subpassage Descr.Ark is present in Len2 (after Gods.Order) and absent from Len1 is described by:


Or to take a more current example, a word xyz is present in L7, but not in L12:

\[ T:12:14:2:10:3:0 + T:7:15:2:9:5:xyz. \]

This formula covers the notion that in MS7 a word xyz is added with respect to the 3rd word on the 10th line of leaf 14 recto of MS12.

Although one usually only notes down the differences of texts, the notion of equality is indispensable for computer usage. In many cases equality can be expressed implicitly by naming the objects with equal names: for instance in a narrative net one uses the passage name as a name. Two passages are the same passage if they bear the same passage name. But this procedure cannot be followed if one wants to identify two words located somewhere in the texttree. These words could be taken from MSS in different languages, so there could be no strict identity on the character level. In such cases the explicit equality operator is used.  

Here we would like to present a provisory list of relation qualifiers used in the collation process. The category of identity (A/A) is not further qualified since the identity is completely determined by the nature of the fundamentals. In every other category we need extra qualifiers to denote explicitly the relation we are aiming at. Since the relations from the categories 0/A and A/0 are mutually inversed, we need to discuss two categories only: omission (0/A) and change (A/B). It is important to keep in mind that the qualifiers are not limited to one research level (for instance the word, or the sentence level), they have their counterparts on the higher levels. In fact a stemma consists of specific genetic relations defined on MSS taken as complete entities.

We describe the relation qualifiers by adding specific letters or codes to the basic category symbols.

A list of relation qualifiers used in text collation:

- + addition without qualification
- - omission without qualification
-S text scratched out
Figure 9. The story of Noah, substory 'Entering'


Figure 9 A. Rabghuzi, Qisas al-Anbiya’. MS Leningrad, Publicnaja Biblioteka, T.H.C. 71, f. 14r.

Figure 9. The story of Noah, substory ‘Entering’

IN SEARCH OF FORMAL IDENTITY AND DIFFERENCE

text overwritten with +W

addition of paraphrase: or more or less redundant extension of a segment like the rephrasing of an already made statement, a pleonasm etc. +P

functional repetition: a strophe, a recurrent rhyme +F

editorial addition (lower case symbols can be used for more distinctions here) +E

material damage -m
dittography: unfunctional repetition of a segment +d

illegible text -i

With respect to change:

paraphrase instead. For instance the reference text (f1) gives a quotation in Arabic, the other text (f2) only a paraphrase in Turkic. *P

synonym instead +S

antonym instead +A

archaic, dialect (lower case symbols can be used for more distinctions here) +D

other language instead +O

change in orthography, spelling, diacritics etc. +E

enclisis +B

binding error

change in writing style (ductus etc.) +W

abbreviation instead +Ab

metathesis +M

imposition: a wrongly interchanging of segments resulting in an incorrect segmental sequence. A typical printing error. +I

reinception: a segment is found in a place where it obviously does not fit nor belong. Possible explanations of this phenomenon can be found in the process of writing (upon the different folia of the diplomata of quaterns), it could be the subsequent result of a binding error of the original or simply the result of some unwittingly made mistake (incorrect reminiscence of the copyist). +R

other syntax instead +T

substitution in a general sense7

Of course this list is not meant to be exhaustive. We only want to show how it is possible to formulate precisely and coherently the most important relations met during collation8.

If we put all this into an example taken from the studied MSS, we get the following results. We have taken the subpassage of the Execution of Gods Order. This subpassage occurs in four MSS:

Len1 سَوَى يُزِينَهُونَ ٱلْقَامُعَاتِ يَغْهَلُونَ. تَغْمِنَ بِرِيْرِنِّد َبِر ٌرِيْهِ كَتَبُوهُ ٍكْنِيْجِ كَفْرُدَيْلِ (f.14r14)

Len2 نَعَٰلَ (A.S.) تَغْمِنُ ٍكَانُوارِلَيْدَنَ (f.21v10) ٍكَفْتُ كَفْتُ بِر يِرِكِ بِر اِرْشِ رِبِنْكِ كَنِمِرِجِ سَلَحَ. ٍقَرَٰثُهُ ٍتاَّلْغَ: ٌفَٰشْلُكْ ٍفَٰشْلُكْ ٍفِحَّةُ مِنَ (f.21v11) ٍكَلِّلَ زَغَٰفَٰنَى اِثْنَانِيْنَ وَاِخْلَاقَٰا. ٍ$
One can assemble all material seen from a specific viewpoint (the reference text; this can be a reconstruction or a really existing one) in a dynamical way. One can formulate hypotheses about affiliations or best paradigmata and test them on the texttree. This will greatly enhance the quality of the proposed best, original or richest text. We do not pretend to have found the final formal solution to all basic collation problems. In fact we do not believe that such a final solution is possible. But still we hold that our approach leads to the construction of a programmed system that possesses the following properties:

- being functionally equivalent to a joint facsimile of all used MSS, by means of a transcription system as sophisticated as the material texts themselves. Of course this is not a real facsimile. It is possible to produce a photographic copy of the text on compact disk in conjunction with the computer representation.
- the text corpus can be studied from any angle and on any level within the accepted criteria.
- the corpus can always be expanded by adding the latest found material or adding new interpretations (nets).
- the treatment of variants is not limited to the perspective of one single reference text. One can shift from one viewpoint to another.
- entities once added can be corrected, reinstored and related to other entities.
- the supported research is not limited to a specific genre.
- objective in the sense that hypotheses can be checked in a clear intersubjective way.

**SUMMARY**

Computer processing of texts and scientific collating of texts (building up an apparatus criticus) using a computer in particular presupposes a thoroughly formal definition and description of the objects and procedures involved. In this contribution we propose a method to tackle the problem. We introduce the notion of a texttree (a formal description of the text material) and of nets ‘thrown’ over the texttree. We also discuss more specific notions involved in the process of text collating, such as the basic categories of text comparison (identity, addition, omission and change) and methods to formally describe notions such as position, relation and qualification. We use the opposition paradigm/syntagma to unify our approach. Some examples are given.