Thomas Jügel

Peculiarities of Avestan Manuscripts for Computational Linguistics[1]

Abstract

This paper will discuss several computational tools for creating a stemma of Avestan manuscripts, such as: a letter similarity matrix, a morphological expander, and co-occurrence networks. After a short introduction to Avestan and Avestan manuscripts and a representation of Avestan peculiarities concerning the creation of stemmata, the operatability of the above-mentioned tools for this text corpus will be discussed. Finally, I will give a brief outlook on the complexity of a database structure for Avestan texts.

Introduction

The Avesta, represented by the edition of GELDNER (1886-96), appears to be a sort of Bible containing several books or chapters, cf. SKJÆRVØ's "sacred book of the Zoroastrians" (2009: 44); and, indeed, in Middle Iranian times (i.e., before 600 AD) there existed a kind of text corpus, rather than 'a book', of holy texts (CANTERA 2004). However, GELDNER's edition disguises the actual texts of the manuscripts because what we have today is not a book but a collection of ceremonies attested in various manuscripts.

Avestan is the term for an Old Iranian language, as such a member of the Indo-European language family. The actual name of the language is not known to us. The name 'Avestan' is taken from Middle Persian texts which refer to their religious text corpus as the "abestā(g)". When manuscripts containing these religious texts came to light for European research, they were referred to as "Avesta" and the language as "Avestan".[2]

Avestan is known to us in two varieties, called "Old Avestan" and "Young Avestan". This is so because they display two different chronological layers of Avestan. However, they also differ in some linguistic respect so that they represent two different dialects of the same language (e.g., genitive singular of *xratu-* "wisdom" is *xratāuš* in Old Avestan but *xraϑβō* in Young Avestan, for further examples see DE VAAN 2003: 8ff.).

The Avestan manuscripts (henceforth MS) can be sorted into several groups, the main grouping is: 1) the 'Pahlavi-MSs', and 2) the 'Sade-MSs'. The Pahlavi-MSs contain the Avestan text plus its translation and commentaries, generally Middle Persian, but there are translations into Sanskrit, Gujarati and/or New Persian as well.[3] The Sade-MSs (i.e., the "pure" MS) only contain ritual instructions in Middle Persian, etc., besides the Avestan text. The Pahlavi-MS served as exegetical texts written for scholarly use only. On the contrary, the Sade-MSs were for the daily use in the ceremonies. These different purposes had an influence on the copying process (cf. Section 1).

The aforementioned grouping can be made by first glance at the MS because of the various writings these MSs do or do not contain. Besides the grouping into Pahlavi- and Sade-MSs, the MSs are further classified into different ceremonies. There are four of them: the Yasna Rapihwin, Vīsprad, Yašt, and Vīdēvdād ceremony. Depending on the season or on the deity who is invoked, there are further differences in what is otherwise the same

ceremony.[4] These latter groupings are veiled in GELDNER's edition and this may lead to wrong approaches when it comes to generating the stemmata[5] of Avestan MSs (cf. Section 2).

1 The copying process

If the scribe copies a MS used for the scholarly work on Avestan with all its commentaries (i.e., a Pahlavi-MS), the main interest is to produce the exact copy of the original. Together with the MS itself, the colophon is also copied, since it serves as a kind of proof of quality when the list of authorities is given. In this process the original was usually not corrected, probably not even read (if the scribe could read Avestan at all). So loss of lines, even of pages, often went unnoticed (cf. CANTERA 2010). Furthermore, it might well be the case that the scribe has mixed different styles of writing. There are, e.g., two versions of the characters ⟨ą⟩, i.e., nasal /a/, and of ⟨y⟩.[6] The one is typical for Iranian MSs, the other one for Indian MSs. We know that Iranian MSs were brought to India. An Indian scribe copying an Iranian MS would have had the choice of copying not just the text but also the style of writing or of converting the Iranian features into Indian ones. This transfer would surely not take place consistently; and, indeed, some MSs show both features.

The scribe who is copying a Sade-MS, which is used in everyday life, would want to produce the best text, not the best copy. As there might have been scribes who could not read Avestan and Middle Persian very well, others were surely experts in it, having a high knowledge of the Avestan ceremonies as well. In the tradition of these MSs, loss of text was usually noticed and the text restored – not always with the correct result as we can say today. A telling disimprovement on the word level was presented by CANTERA on occasion of the conference "Poets, priests, scribes and librarians: the transmission of the holy wisdom of Zoroastrianism" (Salamanca 2009):

> Frequently a final *-əṇg* appears in the manuscripts as *-əṇga*, clearly a reflection of the pronunciation with a final epenthetic vowel. Since this error was known to the priests, they sometimes made hypercorrections. Thus the well-known Indian scribe Dārāb Hīrā "corrected" the right *vīspāṇg. āiiōi* in Y31.2 into the wrong *vīspāṇg. ẏōi*. He obviously thought that *ā* was an epenthetic vowel in the pronunciation of final *-əṇg*. Such hypercorrections occur often.

Here, the scribe took the prefix of the verb *āiiōi* "I ask for" as the epenthetic final vowel of the preceding word. That it was written separately does not raise objections to this wrong analysis since this would be normal for enclitics. However, the reanalysis goes further. The remaining °*iiōi* was understood as the relative pronoun *yōi* (nominative plural masculinum), and ⟨ii⟩ is the orthography of non-initial /y/. Thence °*iiōi* was changed into *yōi*. Another of such erroneous reanalyses is *aēšəm. mahiiā* in Yasna 48.12. The genitive singular of *aēšəma-* "fury", i.e. *aēšəmahiiā*, was split up into two words: *aēša-* "capable" and the genitive singular of the possessive pronoun *ma-* "my". This reanalysis seems to be very old because all of the MSs known to us show some variation at this point.[7]

Besides such corrections, there also occurred alternations of the ceremony, whether because the scribe was following a different custom, or because he had heard of a variant

that he considered better because it was being propagated by a high authority (*mobile variants*).

Obviously, the tradition of Pahlavi-MSs and Sade-MSs proceeded quite differently. We do not expect the same peculiarities of copying processes for both of these groups. The *mobile variants* also blur the border of the various types of ceremonies. Hence, a noteworthy alternation might not be due to the manuscript's belonging to the same group, but rather to external influence on the copying process.

2 The difficulties in generating a stemma for Avestan manuscripts

A huge part of the Avestan corpus is lost. We know this because there are references in the Middle Persian Zoroastrian literature to Avestan passages which were not passed down. Furthermore, the majority of copies is not in our reach – either because they were lost, or because their whereabouts are unknown to the scientific world. There must have been a time when plenty of copies were produced in a year. Some colophons were written by one hand including the year, but the name of the copyist was added later by another hand (CANTERA 2012: 298). There were families whose profession seemed to have been the production of Avestan manuscripts. The ADA project has located more than 300 MSs so far. So, we can consider ourselves extremely lucky whenever we find the rare case of having both the mother MS and its daughter MS or the direct siblings of one mother MS. Such cases show, by the way, that the differences of copies by one and the same author can be much higher than differences of copies from a different copyist (as is the case with the MSs K1 and L4, cf. CANTERA 2012: 329). That is, some copyists did not work very accurately.

Apart from the scarcity of the remaining MSs, we have to consider the impact of the copying process, as described in Section 1, such as deliberate emendations or *mobile variants*. Hence, concentrating on variants which manuscripts may have in common can lead to a distorted picture as it is the case in GELDNER's prolegomena.

The relationship of manuscripts is not necessarily the same as the relationship of the text/textual variants. Manuscripts can be dated according to colophons or by analysing the material (paper, ink). A single MS can be split up into chronological layers when there are emendations and additions of a second hand, which may reveal the influence of another *vorlage*. The text, however, is more abstract. It is *a priori* not clear whether the text of a MS of the 18th century is indeed younger than the text contained in a MS of the 16th century, as aptly put by MINK (2004: 24, italics original): "*the text is the witness, not the manuscript*".

The "Coherence-Based Genealogical Method" (CBGM)[8] combines computational means with philological know-how. At first, the comparison of MSs leads to a so-called "pre-genealogical coherence". A high similarity between manuscripts speaks in favour of a close relationship. Then, for each significant variation a local stemma is philologically stipulated resulting in a textual flow, i.e., each MS is put in chronological relation with the others. An arrow between two MSs in the textual flow does not mean that MS *b* was copied from MS *a*, rather that text *b* is in the textual flow a younger witness being somehow influenced by MS *a*. Furthermore, in a given textual flow a MS can have more than one arrow pointing to it, since its text may have been influenced by the one of several MSs (e.g., in cases of *mobile variants* or collocations from various manuscripts). In order to represent the

various degrees of influence the arrows show different widths, i.e., they are substituted by vectors (CANTERA 2012: 320). Combining the degree of pre-genealogical coherence with the textual flow and the local stemmata yields a global stemma (or stemmata if several sub-stemmata are equally possible), cf. the example in Figure 1 and its discussion thereafter.

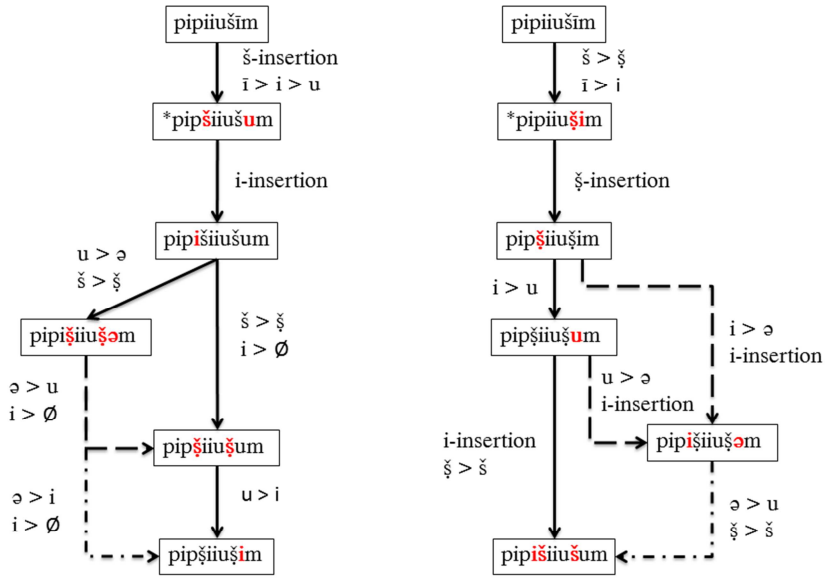

Figure 1: Two possibilities of building a local stemma (data taken from CANTERA 2012: 341)

The original form is *pipiiūšīm* "swollen", a feminine *i*-stem in the accusative singular. In the MSs appear several variants: A) *pipišiiušum*, B) *pipişiiuşəm*, C) *pipşiiuşum*, D) *pipşiiuşim*. There are two equally likely possibilities of how the variants could be arranged in a chain of derivation.

The various ⟨š⟩ characters were confused in the MSs (cf. Section 3.1) so that we may stipulate an intermediary form **pipiiūşīm* (right branch): another ⟨ş⟩ was added by mistake. The length of /i/ and /u/ are not always kept distinct so that the D-variant *pipşiiuşim* may readily evolve. The letters representing ⟨i⟩ and ⟨u⟩ can easily be confused due to their similarity in form, so that we get the C-variant *pipşiiuşum*. In order to explain the B-variant *pipişiiuşəm*, we apply an orthographic rule, viz., if a consonant is followed by ⟨i⟩ or ⟨ii⟩, another ⟨i⟩ is written in front of it (presumably indicating the consonant's palatal pronunciation). The /ə/ might be a phonological reduction of either /i/ or /u/, respectively, revealing an influence of the scribe's pronunciation. For the A-variant *pipišiiušum*, the insertion of an

orthographically motivated ⟨i⟩ is stipulated and, again, confusion of the various characters of ⟨š⟩. A derivation of the A-variant from the B-variant is less likely because /ə/ does not easily change into /u/ (though one could imagine assimilation to the labial in the penultimate syllable). The left branch is equally possible with the same explanations just differently ordered. While in the right branch ⟨š⟩ changes via ⟨š⟩ back to ⟨š⟩, in the left branch ⟨ī⟩ changes via ⟨i⟩ and ⟨u⟩ back to ⟨i⟩.

If we take the textual flow into account, we see that the MSs showing the A-variant are generally prior to those containing the B-variant, which are prior to the C- and D-variants, i.e., the left branch of the local stemma is probably the correct one.[9]

Applying CBGM to Avestan is extremely labour-intensive. One has to digitize manuscripts and to detect and evaluate variants. A high degree of philological knowledge is vital for the evaluation. In order to accomplish such an ambitious task, several scientists of European institutions have agreed on a cooperation which celebrated its constitution as "Corpus Avesticum" on occasion of a workshop held in Frankfurt am Main in November 2011.[10] The work of the philologist can be facilitated by means of computational devices. The following sections discuss their pros and cons.

3 The Avestan Language and Computational Devices

3.1 Simulation of the copying process – interchangeability of characters

In order to set up a local stemma, words at a variant position are aligned in such a way that characters in corresponding positions in sample form a pair of characters. For instance, assuming that the words *aiiese* and *aiesē* would occur as possible variants, the first pair of characters would be "*a-a*". If the variants differ in the number of letters, then a gap is inserted into the shorter word and aligned with the corresponding letter of the longer one (e.g., *a-a, i-i, i-Ø, e-e, s-s, e-ē*). A distance function sums up the distance values of each pair of characters of a variant word pair, normalizes them by dividing by the mean length of the two words, and returning this as an overall value of their distances. With an alignment done by established measures such as the Levenshtein distance (LEVENSHTEIN 1966), each difference in characters would have the same weight.

However, scribes did not randomly substitute one character by another (e.g., writing ⟨k⟩ instead of ⟨a⟩), but rather they were following certain logical rules. Either characters could easily be misread (e.g., ⟨ī⟩ and ⟨ū⟩ in Avestan script), or, according to their phonological surroundings, sounds could be confused and, hence, the characters which represent these sounds (e.g., /a/ and /e/ in a palatal context despite the shapes of these two characters being otherwise clearly distinct). In order to be able to tell trivial changes from non-trivial ones, an Avestan-specific two-dimensional matrix of the interchangeability of characters had to be stipulated (Figure 2 below).[11] The differences of the pair of characters are weighted by applying this lookup table, or matrix of distances, by a distance function. In Figure 2, one dimension characterizes the likelihood of two characters being exchanged due to phonological (and rarely also orthographic) reasons as likely (green), possible (white), unlikely (red). For example, in Old Avestan, final vowels were always long, while in Young Avestan they were always short (with the exception of monosyllabic words and a few flectional endings).

This was, of course, apparent to the scribes as well, who may have tried to archaize Young Avestan texts. So the difference of final *-a* and final *-ā* may simply be of no importance. In palatal context vowels may have been palatalized so that in a sequence like *-iiami-* the variant *-iiemi-* is of little significance.[12] The difference of the sound represented by the characters ⟨β⟩ and ⟨uu⟩ (i.e., bilabial /w/) is a phonetic one, not a phonemic one. The word *ϑβaršta* may also appear as *ϑuuaršta*.

a	**ā**	**å**	**å̄**	**ą**	**ą̇**	**ə**	**ə̄**	**e**	**ē**	**o**	**ō**	**i**	**ī**	**u**	**ū**	
x	9	8	7	1	1	1	1	1	1	1	1	3	1	1	2	**a**
	x	4	8	1	1	1	1	1	1	1	1	3	1	1	1	**ā**
		x	9	1	1	6	5	1	1	1	1	3	1	1	1	**å**
			x	1	1	6	5	1	1	1	1	3	1	1	1	**å̄**
				x	9	3	2	1	1	1	1	1	1	5	1	**ą**
					x	1	1	1	1	1	1	1	1	1	1	**ą̇**
						x	9	1	1	1	1	1	1	4	1	**ə**
							x	1	1	1	1	1	1	2	1	**ə̄**
								x	8	1	1	1	1	1	1	**e**
									x	1	1	1	1	1	1	**ē**
										x	9	1	1	1	1	**o**
											x	1	1	1	1	**ō**
												x	9	8	3	**i**
													x	2	9	**ī**
														x	8	**u**
															x	**ū**

Figure 2: matrix of the interchangeability of vowels

The other dimension of the matrix represents the grade of similarity of the shapes of characters. The higher the figure (1-9), the more similar the characters. The characters ⟨δ⟩ and ⟨γ⟩, ⟨y⟩ and ⟨š̄⟩, or ⟨ī⟩ and ⟨ū⟩ can easily be confounded, though linguistically it is rather unlikely. Aside from the comparison of single characters, character groups also have to be compared. There is a high similarity of ⟨ai⟩ and ⟨ā⟩, ⟨šk⟩ and ⟨š̄⟩, or of ⟨an⟩ and ⟨x^{v}⟩, etc. Phonetically, the difference of ⟨ŋuh⟩ and ⟨ŋ^{v}h⟩ is lacking since both are just two different ways of expressing the same sound: a labialized laryngeal with a nasalizing effect on the preceding vowel (cf. HOFFMANN/FORSSMAN 2004: 45).

There is one striking instance of interchangeability that is not due to the high similarity of the shapes of the characters or of the sounds the characters represent but rather to orthographic conventions.[13] This concerns the characters ⟨h⟩, ⟨s⟩, and ⟨ϑ⟩. All three sounds these characters represent existed in the Old Iranian languages Avestan and Old Persian. In Middle Persian, however, /ϑ/ changed to /h/. In the cryptic orthography of Middle Persian, an /h/ could be indicated by the characters ⟨h⟩, ⟨s⟩, and ⟨t⟩: ⟨h⟩ for wherever it is the normal

representation of /h/, ⟨s⟩ in non-Persian words wherever non-Persian /s/ equalled Persian /h/ due to the results of sound change, and ⟨t⟩ wherever it was the archaic representation of /θ/, which later became /h/. Given all this, when native speakers of Middle Persian pronounced Avestan, they could have substituted /θ/ by /h/ and written it accordingly, or they might simply just confused the three characters ⟨h⟩, ⟨s⟩, and ⟨θ⟩ due to Middle Persian orthographic conventions. In fact, there are only few instances of ⟨s⟩/⟨θ⟩ confusion.[14]

A sporadic variant that is due to the confusion of characters of high similarity or of similar sounds is not significant for the grouping of MSs. Such confusion could have happened at any time. However, if there is a high regularity of such unspecified changes, they might be telling nevertheless.

A programme able to apply the matrix described above can produce a distribution of weighted letter substitutions and will help the philologist to concentrate on relevant variants that allow the stipulation of a local stemma. Those variants that are due to trivial changes will only be evaluated when they are needed for the constitution of a local stemma that comprises significant changes.[15]

3.2 Morphological expansion

The automatic generation of paradigms is helpful for text technological analyses of word forms (e.g., POS) that have not been entered in the digitized lexicon. Therefore, it is necessary to feed the programme all information needed, e.g., sets of endings, inflectional classes, stem alternations, etc. In highly standardized languages such a task can be accomplished with reasonable effort. In Avestan, however, it is much more complicated. To begin with, there is no standardized orthography; the orthographical conventions are rather tendencies. So the rules for the interchangeability of characters described in Section 3.1 have to be applied to the analysis of word forms as well. Then, most of the nominal suffixes have to be entered in as two variants because there is a differing output of endings in so-called *sandhi* context: while, e.g., word final **-as* (ending of the nominative singular masculinum) developed via **-ah* to *-ō* in Avestan, **-as* followed by the enclitic *-ča* "and", i.e., in sandhi context, was preserved (so there is *haomō* besides *haomasča*). Strictly speaking, the paradigm of each declension should show each ending in pausa as well as in sandhi context.

The combination of suffixes may also lead to a different phonological output. So, one cannot simply combine them. For example, the suffix **-ant-* has two different outputs: 1) *-əṇt-*, 2) *-ąt-*. Whether the one or the other output is to be expected depends on the phonological surrounding, i.e., which suffix is following.[16]

Sometimes it is hard to tell whether the variants shown by the MSs are linguistic variants due to dialectal or chronological differences, or whether they are the result of the copying process. However, even in the cases where we do know the regularities, they are so numerous that the task of installing grammatical rules for automatic generation seems hardly worth the effort. As an example I shall explain the paradigm of *pitar-* "father" in detail:

- nominative singular: *ptā* besides *tā* and *pita*
- accusative singular: *patarəm* besides *pitarəm*
- dative singular: *fəδrōi* besides *piθrĕ*

Seeing this irregular paradigm, we may say that "luckily" not more forms are attested. The Indo-European nominative singular **ph$_2$térs* developed as follows: The laryngeal **h$_2$* was either lost or vocalized to *i*. The vowel **e* regularly changed to *a*, and the auslaut **rs* was assimilated to **r* plus compensatory lengthening of the vowel (Szemerenyi's law), followed by a yet unexplained loss of the final **r*. This yields the Iranian output *pitā*, or *ptā*, respectively. The uncommon onset *pt* was simplified to *t*, i.e., *ptā* > *tā*. The Young Avestan form shows the shortening of final vowels, hence *pita*. The Indo-European accusative **ph$_2$térm* developed via **p(i)taraəm* to either *pitarəm* or *ptarəm*, where an anaptyctic /a/ was inserted in the later tradition of Avestan. The Indo-European dative **ph$_2$tréi̯* developed via **fϑrai̯* to *fδrōi*, which displays the Old Avestan development of **ai̯* to *ōi* at the end of a word. Again, an anaptyctic vowel /ə/ was introduced. Besides these irregular forms, there was an analogically introduced stem *piϑra-*, which displays the fricativization of preconsonantal voiceless stops (i.e., **tr* > *ϑr*). The form *piϑrĕ̄* shows the Young Avestan output of word final **ai̯*.

So what synchronically seems like a nightmare for every child or non-native speaker learning this language, can easily be explained by the linguist from a diachronic point of view. The rules, however, do not outweigh the irregularities that are due to phonological effects (sound change), analogical formations (morphological effect), or to reflexes of spoken language (assimilation in the course of recitations).

When it comes to a language like Avestan with such a small corpus of less than 12920 words (DOCTOR 2004: 5),[17] it is easier to annotate every single form by hand. The automatic production of non-attested word forms would always remain highly hypothetical and offer very little insight. Nevertheless, a morphological expander is helpful in suggesting to the philologist the most likely form.

The irregularity not only affects declension or conjugation but also the stems themselves, i.e., not only the grammar but the lexicon as well. The word *napāt-* "grandson" (cognate to English 'nephew') is attested with three different stems: *napāt-*, *naptar-*, and *napa-*. These alternations are not simple mistakes. They are of high interest and show some linguistically well-known patterns. *napāt-* is the inherited form. *naptar-* is a transmutation in analogy to other words denoting family terms like *pitar-* "father", *brātar-* "brother", *mātar-* "mother", etc. This change is due to the semantic class of family terms, most stems of which end in *°tar-*, hence *napāt-* > *naptar-*. The stem *napa-* is based on a regularization process. The many declensional classes of Old Iranian were simplified to a few, the most dominant one being the *a*-declination. Words were extended by *-a-* to make them fit into this class, e.g., *n*-stem *zruuan-* "time" besides the newly formed *a*-stem *zruuāna-*. In the case of *napāt-* the form was shortened to *napa-*. The patterns described are not a sign of degeneration but of language development along the lines of logical reasoning. So, we do not want to emend this. We want to find it. Therefore, each variant gets its own entry in the lexicon representing an inflexion class of the lemma (cf. LINDE this volume).

3.3 Co-occurrences and citation

A handy tool of digitized corpora is a co-occurrence analyzer to check the contexts of a word, i.e., it yields those words occurring with the target most frequently. Therefore, a window is defined comprising the given word and its neighbours. In manuscripts which

contain interpunctuation marking clause boundaries the frame usually is the sentence. Where one cannot detect such boundaries, a sequence of words containing the target is taken instead. In languages rich in morphology, this sequence may be smaller than in those with an analytic system. For instance, German only exhibits four cases (nominative, accusative, genitive, dative) the marking of which are many times syncretistic, i.e., the same suffix is used for different cases. Several nouns only distinguish number (singular, plural) and are not marked for case at all (e.g., *Frau* "woman", *Frauen* "women"). It is the preceding article that makes case forms distinguishable (e.g., *die Frauen* nominative/accusative, *der Frauen* genitive, *den Frauen* dative, all plural). Hence, such languages like German display a huge amount of functional words (articles, auxiliaries, adpositions, and particles). Languages with a rich inflectional system like Avestan do not need these functional words. Instead, they show longer words exhibiting all kinds of functional information in the affixes.[18]

When the window is defined and a query is made, the result will show the word's significant co-occurrences, which can be filtered by their part-of-speech. Such co-occurrences represent valuable information for historical semantics.[19] Tools such as Linguistic Networks[20] allow the visual representation of co-occurrence networks, i.e., not only the target and its co-occurrences are listed but the co-occurrences of the latter ones as well (cf. the following screenshot, Figure 3).

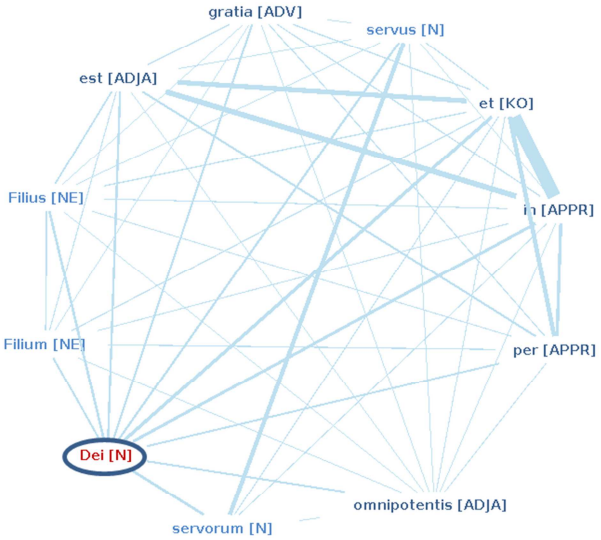

Figure 3: co-occurrences of Latin *dei* "of god"

For Avestan studies, such a query could reveal a differing usage of words in Old and Young Avestan. If Middle Persian is taken into account as well, an alteration of concepts may become visible.[21] For instance, the word *daēnā-* means "religion" in New and Middle Persian (*dēn*, or *dīn*, respectively). However, in Middle Persian another meaning is still detectable. It is the personification of the good or bad deeds of a human. If a man was good, he could expect to meet a beautiful girl in the afterlife who accompanies him to paradise. If he was evil, an ugly, stinking old woman would await him. In the Avestan ceremonies the priest may have contact to the transcendent world and meet his *daēnā-*. The original meaning of *daēnā-* in Avestan is considered to be "view, conception", and not "religion", which has become the traditional translation. Another word of interest is *frauuaši-* "choice", later a personification of the good choices of the ancestors, a guarding spirit.

Avestan texts are the holy texts of Zoroastrianism. However, it is the Middle Persian corpus that is the biggest among the Zoroastrian library. Indeed, we have more texts on Zoroastrianism than holy Zoroastrian texts. Middle Persian texts reveal that there once was a so-called 'Great Avesta' with Middle Persian translation. The Avestan ceremonies we know of today were not necessarily part of this Great Avesta. They may be the textualisation of the spoken ceremonies, i.e., of the practice. Having said this, scientists think that some Middle Persian translations were nevertheless taken from the Great Avesta because they differ in style and translation technique from those that were probably translated directly from the textualised ceremonies. There are few texts that are said to have been part of the Great Avesta (e.g., the Nērangestān, a Middle Persian text with Avestan quotations). If we link the Avestan words, phrases, and clauses with their Middle Persian counterparts, queries will allow classifying and sorting translation techniques, which may differ from text to text. With this knowledge it is then possible to detect Avestan *vorlages* of Middle Persian texts, the Avestan original of which is not known to us. The picture which emerges from such an investigation will show how far Avestan was known to the Zoroastrians of post-Sasanian Persia, i.e., after the Arabic conquest and the spread of Islam. Furthermore, we will get a glimpse into the literary corpus of Sasanian Persia. Even purely Middle Persian texts such as the Bundahišn, an encyclopaedic work, may be based on Avestan *vorlages*. The Avestan Vīdēvdād comprises a legend on the creation of several countries, quite similar to the style of the Middle Persian Bundahišn (chapter 31). So, how Avestan indeed is the Middle Persian corpus?

3.4 Interdependencies of Avestan texts

This section will deal with the complexity of a database the purpose of which is to represent the entire Avestan corpus. The information contained in Avestan manuscripts is allocated to several interdependent segments. As a basis, we can take the Avestan text. Then there are additions and emendations of the text written in the margins or between the lines. These may result directly from the reading or understanding of the Avestan text(s). In the first case, the comparison is drawn to the text committed to memory, which the copyist uses in daily ceremonies. In the second case, these interpolations may result from the Middle Persian translation, which presents yet another layer. Further translations (like into Gujarati) might be based directly on the Avestan text, but are more likely derived from a Middle Persian trans-

lation. So, we can build a hierarchy of dependency. However, an interpolation can bypass an intermediate level and affect a much lower or higher one, e.g., based on the Sanskrit translation of the Middle Persian translation that is the direct translation of the Avestan text, a copyist may decide to "correct" the Avestan text. Besides translations, we also find commentaries that are definitely based on the understanding of the text. These commentaries show influences of the current *zeitgeist* and may have been reinterpreted quite differently by copyists of later centuries. Such reinterpretations – although not changing the wording of the commentary itself – may have had an effect on translations into other languages or, again, may have lead to interpolations of the Avestan text. Avestan text passages are quoted or referred to in Middle Persian texts (e.g., the *Pursišnīhā* "The catalogue of questions"). Although these relations lead outside of the Avestan text corpus itself, viz., to Middle Persian texts, they may reveal the current understanding of the Avestan text at the time the Middle Persian text was written.

The Avestan text itself may be segmented into the Old and Young Avestan texts. There are references to Old Avestan in Young Avestan, and Young Avestan features appear in Old Avestan text segments. The many repetitions – sometimes with small variations and/or short additions – form another set of segments.

Furthermore, there are different ceremonies that only partly display the same text. Variations that belong to different ceremonies could have been judged by copyist as "better" forms (cf. Section 1 "*mobile variants*").

So, we have several layers, some of them arranged horizontally, others vertically, and still others standing in an interdependent relationship. As a corpus is built up step by step, i.e., layer by layer, the interdependency grows and should be taken into account by tools organizing and evaluating the data.

4 Conclusions

I hope to have shown the peculiarities of a language that is only known by its textual sources. Generally, these observations hold to be true for all languages that have not yet developed a standard written form. Avestan is all the more complicated because its oral tradition (later by non-native speakers), its late textualisation, and its subsequent textual tradition led to many effects for which we first need to determine linguistic relevance. Often, we simply still do not know the correct reading of the original. To overcome this, a stemma has to be established. The international scientific cooperation Corpus Avesticum will apply the CBGM method, which combines computational methodology with philological expertise. Several tools will facilitate the work of the philologist, e.g., a tool for finding significant variants by means of a distance function, including a matrix of character interchangeability.

The confusion regarding the original text has prevented comprehensive syntactic studies so far – a job that can easily be accomplished via a well-organized database. Queries on semantics by means of a co-occurrence analyzer will help to elucidate the meaning of unknown words and the development of concepts. Quotation analyses will help to trace the literary-historical development of Avestan and, especially, Middle Persian. These simple tasks require complex spadework: the linking of Avestan with the secondary languages into which it was translated.

An anticipated hurdle is the development of an interactive database that will be available online. Various subcorpora, e.g., a database of manuscripts including their images and metadata, a concordance of digitized texts of the manuscripts, a collection of edited translations, and a database of quotations all need to be interlinked by means of modules such as attestation, lexicon, and grammar. The user should be able to navigate easily from one corpus to the other, or to call up a visualization illustrating how these are linked. Furthermore, the user should have access to adjust, add, and alter information with real-time effect on the linked modules.

Besides the high linguistic and cultural impact of Avestan and the importance of its understanding, the specific problems that the small Avestan corpus presents may motivate us to develop methods and tools that would be useful for other tasks as well.

---

[1] I would like to thank Prof. ALBERTO CANTERA of the University of Salamanca, who so willingly shared his knowledge of Avestan and Avestan stemmatology.

[2] A more detailed survey is to be found in CANTERA (2004).

[3] Note that Indian scripts are dextrograde (left-to-right), while Iranian scripts (ultimately derived from the Aramaic script) are sinistrograde (right-to-left). When it comes to the use of both on the same piece of paper, the scribe faces the problem of organizing the lines in order to avoid one script overwriting the other. This has been prevented by either leaving the rest of the line blank, by jumping into the next line as soon as the dextrograde script reaches the end of the sinistrograde passage (e.g., MS G10), or by turning the leaf 180°, in order to write the dextrograde script upside-down so that it becomes sinistrograde when turned back (e.g., MS S1).

[4] For a more detailed survey on the various types of Avestan MS see CANTERA (2012).

[5] A stemma is a family tree of manuscripts which shows the relationships of the surviving witnesses of a text. Traditionally, each stemma has at its top one original text version.

[6] For a survey on Avestan characters and their encoding see GIPPERT (this volume).

[7] Cf. GELDNER (1886-96: 171 of the Yasna). I re-checked all MSs available on ADA (http://ada.usal.es/paginas/buscador_obra, 3rd April, 2012). Only G18b is with *aēšəm.ahiiā* very close to the original *aēšəmahiiā*. According to TITUS (http://titus.uni-frankfurt.de/texte/etcs/iran/airan/avesta/yasna/yasnavar/yasna.htm), the same is true for the MSs Br2, Jm2, and Jm3.

[8] The CBGM was developed by MINK (cf. 2004 with further references) and adapted to Avestan by CANTERA (2012).

[9] See CANTERA (2012: 341) for a more detailed discussion of this problem.

[10] http://corpusavesticum.hucompute.org/. Constituting members are affiliated to the Universities of Berlin, Bologna, Frankfurt/Main, Göttingen, London, and Salamanca.

[11] For a discussion on letter identification see MUELLER/WEIDEMANN (2012) with further references. – Since the object of our study lies in the past, experiments for stipulating the matrix were impossible. Instead, I knowingly set up the matrix based on my experience and intuition.

[12] It is not always clear when such a difference is due to the pronunciation (probably even by non-native speakers of Avestan) in the recitation and when such changes are the result of sound changes, i.e., a feature to tell dialects or chronological layers apart. Cf. DE VAAN's

(2003: 266f.) discussion of **-čam*, **-ǰam*, **-čam* > **-čim*, **-ǰim*, **-yim*. DE VAAN postulates an intermediary **-čəm*, etc., which is partially preserved in Old Avestan. The development of *ə* > *i* is considered by him to be an effect of "the post-archetype pronunciation", i.e., not a linguistic feature of the language Avestan itself.

[13] Other orthographic conventions, like Indian ⟨y⟩ for Iranian ⟨ẏ⟩, also fulfil the condition of high phonological similarity (in this case both characters represent the same sound). Hence, *ẏąm* vs. *yąm* do not represent two different words or word forms. They are considered to be not two variants but two readings of one variant (CANTERA 2012: 329).

[14] One such example is *ϑrāzdūm* in Yasna 34 §7, which is represented by *srāzdūm* in the MSs Mf1, K37, and Pd (GELDNER 1886-96: 125 of the Yasna), in Br2 (http://titus.uni-frankfurt.de/texte/etcs/iran/airan/avesta/yasna/yasnavar/yasna.htm), and in ML15284 (http://ada.usal.es/paginas/ver/15806).

[15] Confer HOENEN (forthc.) for a theoretical survey on such a programme.

[16] Confer DE VAAN (2003: 624ff.) for rules. For instance, ⟨a⟩ has 12 different inputs, ⟨ā⟩ 13, ⟨ą⟩ 6, ⟨ə⟩ 7, etc.

[17] The number of 12920 words comprises word forms as well, i.e., not only lemmata. However, since DOCTOR also gives compound components as extra entries, the number should be reduced because the first unit of a compound usually does not represent a part-of-speech in its own right. The form is often the stem, or a specific interfix emerges.

[18] The same holds true for agglutinative languages like, e.g., Turkish. In the phrase *bunu yapabileceğinizi söylediniz* "you said that you will be able to do this", the single word *yapabileceğinizi* consists of the following entities: *yap-* stem "to do" + *-abil-* "to be able" + *-eceğ-* for future reference + *-iniz-* "you" (plural) + *-i* for the accusative. That is, what English renders with seven words (*that you will be able to do*) is expressed by a single one in Turkish. This simple example shows that when it comes to comparing languages with one another, language specific features must be taken into account so that whatever is compared (e.g., word length) is indeed comparable.

[19] A respective study is *Virtus. Zur Semantik eines politischen Konzepts von Augustinus bis Johannes von Salisbury*, by Silke Schwandt (PhD-thesis, Frankfurt am Main 2010).

[20] http://www.hucompute.org/ressourcen/linguistic-networks.

[21] Such a study undertaken with classical philological methods is KÖNIG (2010).