

The Voynich manuscript – informal observations on some linguistic patterns

Stephen Bax
Professor in Applied Linguistics
University of Bedfordshire
Email: *bax.stephen@gmail.com*

Abstract

This paper examines patterns of discourse in the mysterious Voynich manuscript, focusing on folio 25v, and folio 116v, the last page of the manuscript. It is argued from this analysis that the element transcribed as ‘daiin’, the most frequently occurring item in the manuscript as a whole, is in fact a discourse marker separating out sense units, functioning like a comma or the word ‘and’, and analogous to the use of crosses in folio 116v. I then suggest that in other ways also, page 25v appears to resemble a prescription, similar to that on page 116v.

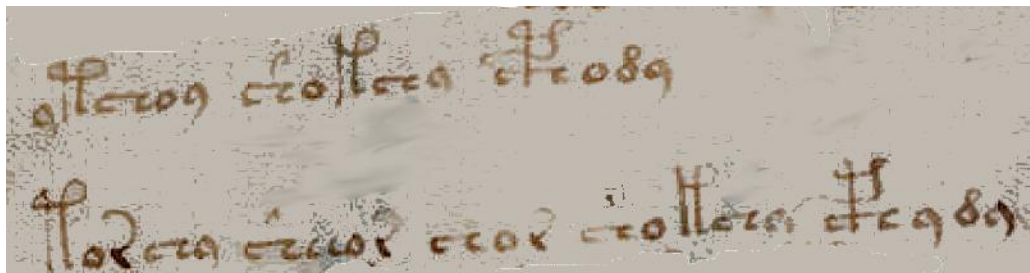
In addition, the paper further argues that on that last page of the manuscript the first of the two Voynichese words (OROR) is probably the name of a plant. It is then suggested that the word is a borrowing of the Semitic word ARAR meaning Juniper, and furthermore that this is the plant represented on f16r. It is suggested that this is the first word to be convincingly translated from the Voynichese text.

Through identifying the function of ‘daiin’, the structure and genre of the text on f25v, and possibly the meaning of the word transcribed as OROR, this paper therefore seeks to offer insights which can help to open up the Voynich manuscript and materially assist in the endeavour to find a complete interpretation of this famously cryptic text.

Numerous attempts have been made to analyse the language and script of the Voynich manuscript (VM) using large-scale statistical analyses of character and word frequencies, drawing on transcriptions created according to various criteria. The aim of this has usually been to compare the statistical frequencies to those of known languages, or similar. By contrast, the analytical approach adopted in this paper comes from a different direction, namely discourse and genre analysis focusing in detail on small parts of the original text to look for linguistic patterns, then attempting to understand what these patterns might signify (see my book *Discourse and Genre*, Bax 2010).

My professional background is in applied linguistics, particularly in discourse and genre, with a specialism in Semitic languages, namely Arabic, Akkadian and a little Hebrew. I mention this because it is useful for the reader to recognize from the outset the biases and limitations which I bring to the

analysis. My reason for writing this paper is that I consider the discourse and genre approach to be bearing fruit, slowly but convincingly, in unlocking some of the many puzzles posed by the manuscript. My starting point was to look for patterns on particular pages. By way of a simple example to illustrate the approach, here is a short such sequence in folio 4v, lines 6-6:



Lines 6 and 7 are transcribed as follows, in the EVA transcription¹:

L6: *ytchoy shokchy cph!ody*

L7: *torchy sheeor chor shokchy cphy!dy*

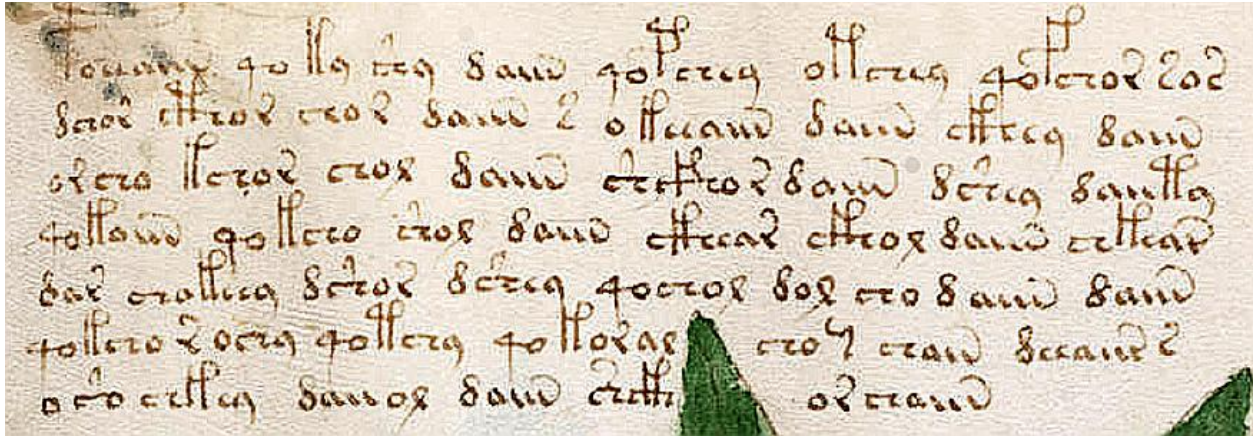
It will be seen that there is a partial repetition here, namely *..choy shokchy cph!ody* in L6 and *chor shokchy cphy!dy* in L7. This kind of pattern is unlikely to be random, and indeed such patterning is normal in natural languages, so the aim of close discourse and textual analysis is to identify what the differences can tell us about the grammar and vocabulary of the language.

In this case there appears to be some sort of inflection, with *ytchoy* changing to *chor* (i.e. losing the prefix plus changing the final character) and *cph!ody* to *cphy!dy* (a change of one medial character). We could surmise that we have here a phrase repeated from the previous line but inflected slightly because of extra words at the start of line 7, or some other reason. The word in the middle, *shokchy*, has not inflected, so it might be a different class of word of a type which does not inflect.

On its own this example can offer limited insight, but this kind of observation, if repeated extensively through the manuscript, could potentially reveal a lot about the (assumed) underlying language. By approaching patterns in the text in this way throughout the manuscript we could start to see bigger regularities, which might then give us clues as to the nature of the underlying language itself.

Having set out my general approach, we can turn to examine in detail a more substantial chunk of text, namely the first paragraph on folio 25v. This text was chosen because it seemed at first sight to exhibit rather unusual patterning, and it is reproduced here:

¹ For an explanation of EVA see <http://www.voynich.nu/extra/eva.html>



It is important to examine the original rather than a transcript so as not to miss some of the key features, but for ease of reference in our discussion, here is the EVA transcription:

- <f25v.P.1;H> poeeaiin.qoky.shy.daiin.qopchey.otchey.qofchor.sos-
- <f25v.P.2;H> dchor.cthor.chor.daiin.s.okeeaiin.daiin.ckhey.daiin-
- <f25v.P.3;H> orcho.kchor.chol.daiin.shcfhor.daiin.dshey.daiity-
- <f25v.P.4;H> qokaiin.qokcho.shol.daiin.ckhear.ckhol.daiin.chkear-
- <f25v.P.5;H> dar.chakeey.dshor.dshey.qochol.dol.cho.daiin.daiin-
- <f25v.P.6;H> qokcho.r.ochy.qotchy.qokoral.cho-!chain.deeaiir.s-
- <f25v.P.7;H> oso.chkey.daii!ol.daiin.shckhy-orchain=

What first caught my eye about this text was an unusual degree and types of repetition across lines. The repetition of the ‘daiin’ word stands out approximately in the centre of the first four lines, as does the use of words starting with ‘ch’ preceding ‘daiin’ each time. The more I looked at the text the more patterns emerged word-by-word and line-by-line, as will be discussed below.

Line	Ch/ Sh element	Daiin	Third element
1	shy	daiin	
2	chor	daiin	
	ckhey	daiin	
3	chol	daiin	
4	shol	daiin	
	ckhol	daiin	
5	cho	daiin	daiin
6	-		
7	chkey.	daii!ol	daiin

TABLE 1

Before discussing these features in greater detail, it is worth asking a question which in my view has not been considered enough when addressing the VM, namely where is the punctuation? The obvious answer is that there is none - but in that case we must ask how the reader could know where the 'sense-units' begin and end? Although historically many scripts had little punctuation, they almost always had instead some form of 'discourse marker' to help the reader to follow the writer's flow of ideas. An example is the Latin suffix '-que' to signify 'and'. Another is the word 'hal' in classical Arabic, an essentially empty word signifying that the following sense unit was to be read as a question. It had no translatable meaning beyond flagging up the function of the sentence as a question, what we term a 'discourse marker', no more.

In my view potential examples of such empty discourse markers in the VM are the initial symbols transcribed as 'p' and 'f', at the start of numerous pages, often decorated. As Currier noted years ago, "[t]hey (p , f) appear 90-95% of the time in the first lines of paragraphs, in some 400 occurrences in one section of the manuscript." (http://www.voynich.nu/extra/curr_main.html). This in itself implies that they are being used to indicate or highlight the first line of a text. More to the point, they occur 107 times as page initial (93 pages with 'p' and 14 with 'f'). Since it is highly unlikely that the author would find actual words beginning with these letters to start these pages, it is highly probable that the symbols are semantically empty markers used simply to flag the start of a page, just as we use a semantically empty full-stop to indicate the end of a sense unit.

It is also apparent that the letters we transcribe in EVA as 't' and 'k' also serve a similar function to signal a new paragraph. On almost every page we see the 'p' as a page starter and then either a 't' or a 'k' starting later paragraphs. This again cannot be coincidence, nor is it likely that the writers found words beginning with those letters specifically. The most logical deduction is that they are possibly empty discourse markers, prefixed to words, signalling a new paragraph.

This is a possibility to which we will return. Coming back to the analysis of folio 25v, and the discussion and illustration above, my aim was to identify patterns in the text and then to interpret the *function* of those elements. In terms of patterns, two of them stood out in that page most prominently, namely the element 'daiin' repeated not only in the middle of the first four lines, but five more times. Considering the fact that this is the most frequent item in the manuscript as a whole, this frequency was perhaps to be expected, but what is noteworthy here is that it is never inflected in any way, whereas it follows words beginning with 'ch' which apparently did inflect in some way. This can be seen in column TWO above, with 'chor', then 'chol' and so on.

The discourse function of 'daiin' – one possibility

After some examination it struck me that one possible function of 'daiin', so frequent as it was, yet not changing, was as a kind of divider between sense-units, what we could term in technical jargon a 'discourse marker' acting to indicate to the reader the sense break. In plainer language, the possible function of 'daiin' is simple but important – it acts much like the word '*and*', or a modern comma.

‘Daiin’ might have a literal meaning, but if this suggestion is correct, any literal meaning is fundamentally unimportant in functional terms, since it appears that its essential function here could be to show the reader where a small sense-unit ends. In some cases it is doubled (as in folio25v, line 5) probably to signal a more substantial sense-break, more like a full-stop. (This doubling occurs 17 times in the manuscript, with one tripling on folio 89r2.). But I suggest that usually it acts as a discourse marker of continuation, connection or break, as in our ‘*and*’ or comma.

To summarise, the evidence for seeing ‘daiin’ as a discourse marker analogous to ‘*and*’ or comma can be set out as follows:

-Firstly it goes some way to answering the question posed above about how a reader would break up the text in the absence of any other punctuation marks. ‘Daiin’ gives the reader a clear guide as to how to recognise the start and end of short sense-units.

-Secondly, it explains why ‘daiin’ is - by a significant margin - the most common ‘word’ in the whole manuscript ; it is used a lot because there are many sense units to divide, just as the comma and ‘and’ are high frequency items in English.

-Thirdly, ‘daiin’ never occurs at the beginning of a page, as you would expect with something acting as a continuation marker. It does appear at the start of some lines, but that simply means that the sense unit ended with the last word on the line before, and the new one is about to begin.

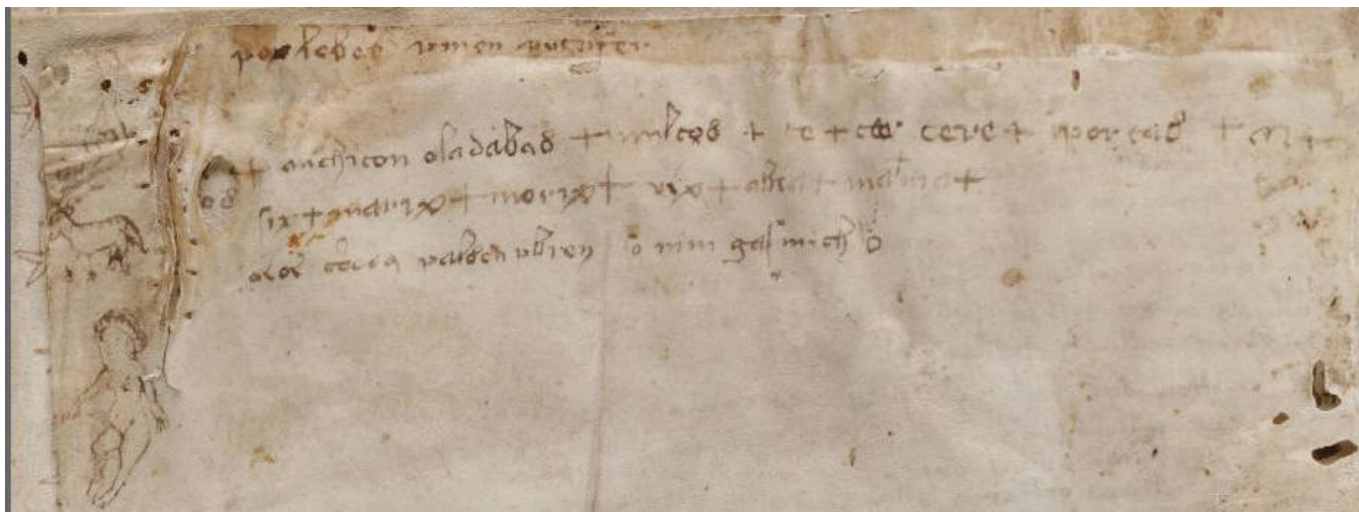
A fourth reason can be found in another part of the VM, namely the last page, to which we can now turn.

The final page – a recipe or prescription?

My interest in the Voynich manuscript began in early 2012, but I was inspired to look at it more closely following my attendance at the Voynich 100 conference in Italy in May 2012. Among the numerous interesting papers given at that event was one presented by Johannes Albus concerning the final page of the manuscript (116 v), in which he argued convincingly that the text is a recipe in Latin and German, with two words in ‘Voynichese’².

Albus’ interpretation appears to me convincing. He explained that the text prescribed a way of using Billy Goat’s liver as a remedy for wet rot, a skin condition, and his analysis was supported by numerous examples from contemporary recipes and other sources, as well as by reference to the picture of the goat and liver in the margin. From this he argued that the text was a ‘recipe’, although I prefer to see it as a ‘prescription’, as Albus’ evidence shows the text to be recommending a mixture for medicinal use, and not merely offering instructions for creating the mixture as in a recipe. I reproduce the original VM page here.

² <http://www.voynich.nu/mon2012/mon07.html#P4>



Albus' transcription and gloss is as follows:

Transcription with abbreviations and omissions in square brackets

- L1 poxleber umen[do] putriter.
- L2 + an[te] chiton olei dabas + multas + t[un]c + t[an]ta[a](?) cer[a]e + portas + M[ixtura] +
- L3 fix[a] + man[nipulis] IX + mor[sulis] IX + vix + alt[e]ra + matura +
- L4 (two ciphered words) pals [ein]en pbrey so nim[m] gei[s]smi[l]ch O

Translation (Johannes Albus)

Billy goat's liver for wet rot
 At the membrane you gave oil, then you bring a lot of the much(?) wax, in a fixed mixture: 9 hands full, 9 morsels (from) the only just double mature
 (two ciphered [Voynichese] words), squash it into a paste, then take goat's milk.

The fact that the text contains two words in 'Voynichese' is significant, since it means that it was not simply a later addendum by an unrelated scribe, but is linked at least tangentially to the rest of the VM. As such it could serve as a help to its interpretation, for reasons we can now consider.

If we examine Albus' interpretation we note that the prescription has a clear structure, starting with the heading on line 1 which indicates the nature of the preparation and also its medicinal use. Line 2 and the

start of line 3 offer an instruction with verbs in the second person, namely ‘dabas’ (imperfect or future of ‘dare’ to give) and ‘portas’ (present of ‘portare’, to carry), although why the tenses are different is unclear. This is followed in line 3 with further ingredients and quantities to be added, with Line 4 offering the two Voynichese words, followed by further instructions in the form verb + noun.

Looking at the two Voynichese words, it appears possible from the structure of the prescription that they might also contain a noun and a verb, given their position in the text. The words have been transliterated as ‘oror sheey’ (Palmer 2004, <http://inamidst.com/voynich/michitonese>).

I propose that the first of these indeed a noun, in fact the name of a plant, for the following reasons. The most obvious reason is that ‘oror’ is the label of part of a plant illustrated on folio 102 v2 Line 1, as follows:



This label makes it highly probable that OROR refers to some sort of plant. If we look at its distribution though the VM, analysis of the word sequence ‘oror’, in isolation or as part of a word, reveals 21 instances in the manuscript as a whole, distributed as follows:

Rank	Item	Frequency
1	oror	5
2	choror	3
3	toror	2
4	doror	1
5	loror	1
6	orory	1
7	sororl	1
8	poror	1
9	sorory	1
10	ooror	1

11	ytorory	1
12	pchoror	1
13	okorory	1
14	pororaiin	1

This frequency count is consistent with ‘oror’ being a noun as opposed to being a more frequent part of speech such as a preposition, with ‘oror’ – the most frequent variant - being the bare form, the initial ‘ch’, ‘d’ and so on being prefixes of various sorts, and the ‘l’, ‘y’, and ‘aiin’ being suffixes.

The three examples beginning with ‘p’ are all paragraph or line initial, i.e. poror (15v P 1), pchoror (f104v P 27), pororaiin (f108v P 27), consistent with the analysis discussed above that ‘p’ functions as an initiating prefix, with little or no additional semantic content.

The most interesting occurrences, however, are in the Herbal section on page 15v, L1, and on page 16r (P2 L10), which are facing pages with two plants illustrated. If the ‘p’ is indeed taken as an empty initiator, then it is possible that, POROR being the first word, it indicates the plant being illustrated. The second example on the same double page spread, TOROR, is also paragraph initial, so if the ‘t’ is again an empty discourse marker signalling the start of the paragraph, it is again possible that OROR refers to the plant illustrated. Indeed with only 21 occurrences of the sequence in the whole manuscript, we would expect one every 5.5 pages on average, so two on the same double page is significant.

I suggest further that the Voynichese ‘OROR’ might represent the word ARAR, which is an Arabic and Hebrew word for *Juniper* or *Juniper Berry*. (Note that the letter A in this transcription from Arabic/Hebrew stands for the semitic guttural consonant AYIN, and not a vowel *per se*). A common variety of the Juniper was the *Juniperus Oxycedrus* plant, with reddish berries and spiky leaves, common throughout the Mediterranean west to the Apennines³ and east to Iran, which was used to make Oil of Cade, an ancient remedy which has been described as follows:

Uses.—Oil of cade has been used locally, by the peasantry, in the treatment of the cutaneous diseases of domestic animals almost from time immemorial. More recently it has been largely employed in the treatment of chronic eczema, psoriasis, and other skin diseases of man...

http://www.henriettesherbal.com/eclectic/usdisp/juniperus-oxyc_oleu.html

This is of interest to us because Oil of Cade’s use as a skin treatment fits well with Albus’ interpretation of the text on VM page 116v as a prescription for wet rot, a skin complaint. In other words in medicinal terms the identification of OROR with juniper fits well with its occurrence in the prescription translated by Albus.

Juniperus Oxycedrus

³ http://www.henriettesherbal.com/eclectic/usdisp/juniperus-oxyc_oleu.html

While the plant on folio 15v looks nothing like any form of Juniper, the plant on the facing page (16r) closely resembles the *Juniperus Oxycedrus* plant, with its distinctive red berries and spiky leaves, as can be seen in the pictures below, of the VM plant on the left and the *Juniperus Oxycedrus* on the right.



Voynich 16r



Images from:

http://en.wikipedia.org/wiki/File:Juniperus_oxycedrus.jpg



<http://www.phrygana.eu/Flora/Cupressaceae/Juniperus-oxycedrus-macrocarpa/Juniperus-oxycedrus-macrocarpa.html>

Pictures of *Juniperus Oxycedrus*, for comparison

The fact that the word *OROR* is mentioned twice on the same double page, both times in paragraph initial position, seems convincing evidence that it is referring to the plant in the picture, whose berries and star-shaped leaves are strikingly similar to the *Juniperus Oxycedrus*.

It is worth noting that the juniper was familiar to 15th century medicine. The medicinal manual entitled the '*Liber medicinarum sive receptorum liber medicinalium*' from around 1475-1500 by John Arderne, in the

special collection of Glasgow University library, illustrates a process of distilling Juniper oil, illustrated below, which shows the importance of the plant in medicinal thinking at the time.



Arderne's illustration of the process for distilling Juniper oil

<http://special.lib.gla.ac.uk/exhibns/month/may2006.html>

For these reasons – drawing on the linguistic, medicinal and pictorial evidence, I suggest that OROR can with some confidence be translated as ARAR, a borrowing from Arabic/Hebrew which is still used today. Indeed the Arar Tree is the national tree of Malta (although not the *Juniperus Oxycedrus*) and the word has a relatively wide and long-established currency.

Folio 25v revisited

The above digression to examine folio 116v and Albus' interpretation of it will now allow us to see other patterns in the text in folio 25v. In the first place, we can note that the text in 116v which Albus was discussing is divided up into sense units separated by a + symbol. These do not divide words, but larger units of meaning, so for example the words in "an[te] chiton olei dabas in line 2 are not each separated by crosses. It is not always clear to the modern reader why the sense units are separated in this text (e.g. why 'multas' and 'tunc' form separate units) but what is clear is that the author considered it important to show those separations with a cross, *in addition to* leaving spaces between each word.

This kind of sense-division on f116v is precisely the same as the function of 'daiin' which I argued for above when discussing f25v. The fact that this same feature occurs in the same Voynich manuscript, in folio 116v, mainly in Latin and German, is further evidence for the hypothesis that the element 'daiin' is operating likewise as a sense-divider, equivalent to the cross on folio 116v and to a comma or 'and' in other manuscripts. Combined with the arguments set out earlier, for example that this interpretation is consistent with 'daiin' being the most common 'word' in the VM, the hypothesis seems to me a strong one, and worth testing in future examination of the manuscript as a whole.

The genre of the text on page 25v

I would further suggest that the text on page 25v, transcribed above, might in fact be a prescription like that which Albus analysed on f 116v. This is still speculative, but it is noteworthy that in format, taking ‘daiin’ as a sense divider, the structure of the text closely resembles the prescription analysed by Albus, as follows:

Possible Structure	Pattern of text in 116v	Possible analysis of f25r as a prescription (with breaks at each occurrence of ‘daiin’)
1. nature of the preparation and its medicinal use.	line 1 .	(Line 1) poeeaiin qoky shy <i>daiin</i>
2. instruction, with verbs in the second person	L2-3	Line 1 cont.) qopchey otchey qofchor sos (Line 2 dchor cthor chor
3. ingredients, in the form of nouns and numbers	L3	s okeeeaiin <i>daiin</i> ckhey <i>daiin</i> orcho kchor chol <i>daiin</i> shcfhor <i>daiin</i> dshey daiiity qokaiin qokcho shol <i>daiin</i> ckhear ckhol <i>daiin</i> chkhear dar chakeey dshor dshey qochol dol cho <i>daiin daiin</i>
4. Further instructions with noun and verbs	L4	qokcho r ochy qotchy qokoral cho-!!chain deeaair s oso chkey daii!ol <i>daiin</i> shckhy orchaiin

Although this is speculative, it is certainly possible that this text is a prescription, with the high incidence of *daiin markers* in the middle of the text indicating different ingredients, mirroring the high number of crosses in the middle of f116v.

Close observation of the original text suggests that the single character which has been transcribed as ‘s’ (in ‘s okeeeaiin’ line 2) does not look like other characters transcribed as ‘s’. but rather resembles the Arabic numeral ‘2’, so it could in fact be a number for a following ingredient. However, this possibility requires more translation of the underlying language in order to evaluate it fully.

Summary

In summary I propose the following:

1. The word transcribed as ‘daiin’ is a discourse marker signalling a sense-break, similar to the English word ‘and’ or a modern comma.

2. The word transcribed as OROR which appears on f 116v, and also significantly as a plant label on f 102 v2 Line 1, and twice on the facing pages 15v and 16r, is probably the name of a plant
3. I suggest that OROR refers to juniper, being a possible borrowing from Arabic/Hebrew ‘ARAR’ and linked to the plant depicted on f 16r. It might well be the *Juniperus Oxycedru* owing to the strikingly similar spiky star-like leaves.
4. The text on f 25r could be a prescription similar to that on page 116v.

Implications

This analysis has a number of implications, which can be set out as follows. In particular I suggest that if this analysis is correct, as the weight of evidence suggest that it might be, then OROR is the first Voynichese word to be interpreted with any confidence. In addition the analysis suggest that

- a) The underlying language is probably a natural one (though it could be encoded);
- b) The script might be at least partly alphabetical rather than fully syllabic, logographic or something else;
- c) The Herbal pages are actually referring to plants such as those depicted – as indeed seems logical; each double page might be discussing *both plants* rather than each being discussed on ‘its own’ page;
- d) Other pages could be prescriptions like 116r.
- e) The manuscript could borrow other words from Arabic/Hebrew. This does not of course mean that the underlying language is necessarily Arabic/Hebrew or anything else, as lexical borrowing is common. However, it does suggest an eastern Mediterranean provenance might be likely.

Conclusion

In my view this approach to analysis is a potentially fruitful one, but there is obviously still a lot of work to be done before the manuscript can yield up its secrets. I would welcome any feedback on any of the ideas presented here.

Stephen Bax, June 2012, revised Nov 2013