Mining the Differences between Penninc and Vostaert

Karina van Dalen-Oskam
(karina.van.dalen@niwi.knaw.nl)
Dept. Dutch Linguistics and Literary Studies
Joris van Zundert
(joris.van.zundert@niwi.knaw.nl)
Dept. Dutch Linguistics and Literary Studies

The Middle Dutch *Roman van Walewein* (Romance of Gauvain, ca. 1260) was written by two authors, Penninc and Vostaert. Only one manuscript containing the complete text, explicitly dated as copied in the year 1350, is left to us. Some fragments of another, probably somewhat younger manuscript contain about 400 lines. The text in the complete manuscript consists of 11,202 lines of rhyming verse. The manuscript was written by two clerks. The first seems to have written the lines 1-5.781 and the second the lines 5,782-11,202.

The second author, Vostaert, explicitly claims to have added about 3,300 lines to Penninc's text. Because scholars of Middle Dutch literature came up with other amounts, we decided to try out modern authorship attribution techniques to find out whether these would point to a specific line in the text where the text before and the text after contrasts most. We used a lexical richness measure, Udney Yule's Characteristic K, and Burrows's Delta, measuring the differences of frequencies of the most frequent words in different parts of the text. We split the text into largely overlapping parts of 2000 lines, moving through the text in order to search for an exact line in the text where the contrast before and after would be the most significant. For measuring Burrows's Delta this meant that for the sake of our focus on one text (or two, in a way), we considered the text as a group of texts' and every part' of 2000 lines as a separate text, to be compared with the other 'texts'.

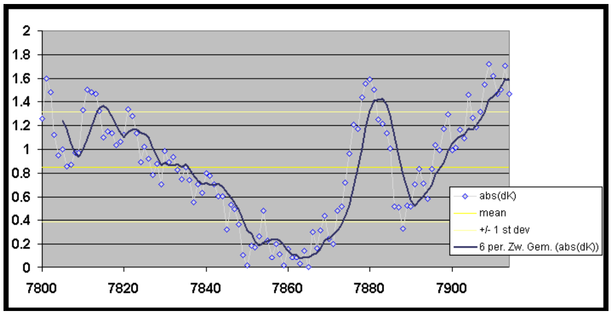

Figure 1: Lexical Richness according to Yule's K.

At the conference in Gothenburg in 2004 we were able to show that both measures yielded the lines 7,881-2 as the point of the most contrast. In Fig. 1 we present the results of Yule's K for that part of the text and in Fig. 2 the results of our creative use of Burrows's Delta can be found. It is very intriguing that both measurements point to the same place in the text. This suggests that line 7,882 could very well be the place where Vostaert took over from Penninc.

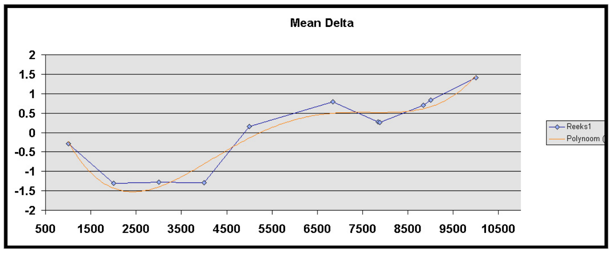

Figure 2: Differences in frequencies of the 150 most frequent words according to Burrows's Delta

We continue our research by concentrating on a quantitative analysis of the differences between the two parts of the text. What are in fact the lexical differences between the text parts before and after line 7,881-2? To find out, we made a list of lemmata (headwords, comprising all spelling variants or inflections etc. of a word) that occur significantly more in the lines before and in the lines after. The top of this list looks as follows:

		stdev	>0.05242999
		mean	0.0166
	Penninc		*z-score*
be, his	zijn	0.8413	15.7293
I	ik	0.8042	15.0217
me	mij	0.6790	12.6328
you	gij	0.5059	9.3325
my, mine	mijn	0.4223	7.7364
may	mogen	0.3158	5.7060
it	het	0.2957	5.3222

stand	staan	0.2665	4.7663
we	wij	0.2514	4.4775
lord	heer	0.2328	4.1224
that	dat	0.2195	3.8692
yonder	gene	0.2137	3.7587
your	uw	0.2131	3.7465
you	u	0.2095	3.6793
say	zeggen	0.2022	3.5387
god	god	0.1903	3.3124
live	leven	0.1774	3.0663
come	komen	0.1702	2.9290
need	moeten	0.1653	2.8359
gate	poort	0.1650	2.8300
see	zien	0.1599	2.7316
squire	knaap	0.1524	2.5898
then	doe	0.1485	2.5157
give	geven	0.1485	2.5150
well, rather	wel	0.1479	2.5043
over	over	0.1474	2.4931
king	koning	0.1454	2.4555
thus	dus	0.1396	2.3445
stay	blijven	0.1392	2.3375
inside	binnen	0.1267	2.0992
not	ne	0.1229	2.0275
at	aan	0.1147	1.8707
shall	zullen	0.1038	1.6623
you	jij	0.1034	1.6550
loyal	trouw	0.1011	1.6111
go	gaan	0.1009	1.6075
serpent	serpent	0.0958	1.5093
allow	laten	0.0954	1.5030
desire	begeren	0.0915	1.4280
day	dag	0.0878	1.3569
where	waar	0.0821	1.2481
all	al	0.0807	1.2211

		stdev	0.03920838
		mean	0.0167
	Vostaert		*z-score*
the, this	die	0.6234	15.4755
he	hij	0.4112	10.0614
to	te	0.3670	8.9353
knight	ridder	0.3659	8.9071
large	groot	0.3406	8.2613
duke	hertog	0.3051	7.3573
very, pain	zeer	0.2951	7.1002
they, she	zij	0.2886	6.9355
Walewein	walewein	0.2823	6.7757
there	daar	0.2748	6.5846
so, thus	zo	0.2260	5.3397
of	van	0.2242	5.2924
Isabele	isabele	0.1844	4.2767
maiden	jonkvrouw	0.1813	4.1977
hit, slay	slaan	0.1607	3.6728
in	in	0.1382	3.0998
horse	hors	0.1349	3.0160
how	hoe	0.1348	3.0117
self	zelf	0.1334	2.9774
other	ander	0.1330	2.9662
fox	vos	0.1228	2.7068
no	geen	0.1196	2.6245
to	toe	0.1171	2.5612
man	man	0.1131	2.4601
many	menig	0.1074	2.3153
black	zwart	0.1023	2.1845
also	ook	0.0985	2.0859
begin	beginnen	0.0980	2.0739
because	want	0.0969	2.0465
brave	stout	0.0961	2.0252
speak	spreken	0.0957	2.0155
to	tot	0.0942	1.9779
helmet	helm	0.0925	1.9352
(some)one	men	0.0918	1.9169
sweet	lief	0.0912	1.9009
on	op	0.0910	1.8953
blood	bloed	0.0884	1.8290
and	en	0.0873	1.8027
walk	lopen	0.0852	1.7485
merciful	goedertieren	0.0820	1.6672

hour	stonde	0.0812	1.6466
do	doen	0.0804	1.6262

[etc.]

Summarizing, Penninc makes significantly more use of the first and second person of the personal pronoun, in contrast to a significantly higher use of the third person by Vostaert. Penninc also applies a lot more modal verbs. But why? Are there several reasons for these differences, or can all be explained by only one or two special effects' of the individual authors?

The first hypothesis we will explore is that a difference in the amount of *dialogue* between the two parts of the text may give rise to several of the differences we have found. The paper will investigate whether this is the case. We will present an analysis of the vocabulary of both authors differentiating between dialogue, narrator's text, and erlebte Rede' (narrated monologue). We will also list other possibly differentiating elements and test whether these play a part in the contrast we discovered by using Yule's K and Burrows's Delta. This qualitative phase in the research is meant to yield an overview of elements contributing to the (quantitative) contrast on the one hand, and to lead us to a list of key elements in the lexicon of the two authors on the other. The list of actual differences will be the input for a new quantitative and qualitative literary analysis of the character and voice of Penninc and Vostaert. Furthermore, we will look forward to the next purely quantitative step we hope to take, in which the results of the above can help us to establish a formula for authorship distinction in the genre of Middle Dutch Arthurian Romance, and help us, so to speak, to leap from the mining to the modelling of the differences.

Bibliography

Burrows, J. "'Delta': a Measure of Stylistic Difference and a Guide to Likely Authorship." *Literary and Linguistic Computing* 17 (2002): 267-287.

Burrows, J. "Questions of Authorship: Attribution and Beyond." *Computers and the Humanities* 37 (2003): 5-32.

Es, G.A. van, ed. *De jeeste van Walewein en het schaakbord van Penninc en Pieter Vostaert.* 2 vols. : Zwolle, 1957.

Holmes, D.I. "Authorship Attribution." *Computers and the Humanities* 28 (1994): 87-106.

Johnson, D.F., and G.H.M. Claassens, eds. *Dutch Romances I: Roman van Walewein.* Trans. D.F. Johnson and G.H.M. Claassens. Cambridge: Cambridge, 2000.

Love, Harold. *Attributing Authorship: An Introduction.* Cambridge: Cambridge, 2002.