COSMOPOLIS

Cosmopolis #11 November, 2000

Contents:

REPORT FROM MILAN

In Milan the VIE held several meetings with Sfera. We got to know the people we will be working with, discussed production, financial and shipping strategies, as well as the construction of the books themselves.

We visited the printing factory to see the machine on which our books will be printed. It uses a photocopy process to print pages from electronic files, gathers and folds the sheets, and sews them into signatures. This machine is the only one of its kind in existence. It was designed, by the printer himself, for runs of from 200 to 1000 sewn books, and includes a module (nuncupatory for the VIE) which binds on a cover.* We also visited the bindery, the oldest in Italy, which prepared a sample VIE volume. It was not possible to prepare a facsimile as we had hoped; this would have meant ordering the exact materials and doing more printing work than was practical. But the sample volume does show us that format, paper, binding style, cover materials and colors do make a fine ensemble.

Particularly gratifying was the excitement expressed by the Italians for the VIE project and the high quality books we want to produce. They gave us the benefit of their experience, and have even become interested in producing an Italian language Vance Edition, a project which must await publication of the VIE itself. This is an encouraging sign that the VIE is already sparking wider interest in Vance among literary circles.

VIE Set Price

Because the price of materials is always changing, it is impossible for Sfera to predict exactly what they will be when the VIE goes to press. However, we can say that the price will be in line with our original projections: $1000 to $1200 for the standard, and $3000 for the deluxe.

The bad news: packing and shipping present more complications than we anticipated. The VIE will occupy almost 4ft. of shelf space and a set might have to be sent in two crates. There are also other complications which have made us realize that, almost certainly, we will have to charge separately for shipping, which means the cost will be as above, plus shipping. There may also be the possibility for people to pick up their sets personally in Milan — a great place to visit by the way. An important concern is reliable delivery of undamaged books, and the capacity to replace damaged copies.

The good news: Not only will the deluxe edition be full leather, but the standard edition will have a leather spine. Our original plan of having sewn books with flexible covers, after discussion with Sfera, now again seems feasible (this is a feature of some of the finest Italian books), and will probably be adopted for the standard edition. The deluxe edition will be hardback, bound in full leather (best quality goat) with gold and red stamping on the spine the covers, gold leaf on the book block, marbled end–papers, and other features still being explored. The standard edition will be lighter in color than the deluxe edition, have gold stamping on the spine, and printing on the covers. Both editions will use a laid paper reserved for the best books, and each volume will probably include a reproduction of an etching or drawing as frontispiece illustration. The VIE edition as a whole, standard and deluxe, will be as durable and handsome as it is possible for books to be, and will grace the shelf of even the most prestigious and exclusive library. The VIE will not only be the Vance edition of reference for all time but guarantee that the ensemble of Vance's work will be preserved for posterity in its ideal form. Part of this guarantee is that VIE subscribers inhabit all four corners of the globe.

When Will The VIE Be Delivered?

We have the price information we need but, since we are not yet as sure as we want to be about our internal

project schedule, we are delaying our call for a down payment. We are currently trying to organize ourselves around a delivery date sometime in the fall of 2002. Assuming we can firm this up in the next month, which at the moment looks possible, we will ask for the down payment in the December *Cosmopolis* (#12). This down payment will be $350 for the standard, $1000 for the deluxe, and subscribers will be given 30 days to send in the money. *THIS IS NOT A CALL FOR A DOWN PAYMENT*. We have been making these noises since August, so this is just another warning that the call is imminent; keep dropping those pennies in the piggy bank! Failure to send in the down payment on time will mean losing one's place on the subscriber list. This is cupatory only for those who signed up early enough to have a slot among the first 200. Late payers will be assigned new places at the end of the list. Modalities of payment are yet to be defined; though PayPal, the Web-based credit card payment method, has recently begun to accept non-US payments. This will, in theory, allow non-US purchasers to use their credit cards to remit their payments.

---

*In the same room was another, much larger, machine which can produce a paper back book with a full-color cover, using no other inputs than a roll of paper, a stack of cover material, and a digital text file. Thanks to this installation the printer can provide very small runs, usually of classics and other books that need to be continually in print. It therefor solves the inventory problem created by the need to lower book production costs by making large runs. This is POD: printing on demand. But this machine is about 80 ft. long, impressive and complex looking, and though maximally automated still requires some manual intervention. The image of POD as refrigerator sized boxes in the back room of every book store which at the press of a button spit out a book while the customer waits, is, therefor, science fiction. Still, it is a promising development in that it has brought the cost of low-quality paperback book production down to where small runs are economically feasible. The great innovation, of course, is digital photocopy printing which allows the printing presses to be sidestepped. However, all the other steps of creating books remain: organizing, cutting, folding and attaching the paper. Large run publications are still more economical using printing presses.

PROOFREADING UPDATE

By the time this article sees print, the proofreading team will have passed the nine-million-word milestone. The end of the first phase of proofreading is in sight. When the assignments you can see on the website: http://www.cs.wisc.edu/~suan/vie/public/StatusByTitle-c.html – have been completed, almost all of our texts will have had at least three proofreads.

As TI gets underway proofing work will change. "Pre-proofing" (meaning pre-TI; what we are doing now) will slow down as texts move into TI and then Composition. As texts roll out of Composition "Post-proofing" will take off. Post-proofing will be even more intensive than Pre-proofing. Note that while Pre-Proofing has a taint of TI to it, Post-Proofing will not. The Post-proofing method will be as follows: 10 PDF (read only) copies of the typeset book will be distributed to 10 proofers who will work on it simultaneously, noting any errors in a report. These reports will be collated and controlled, and any needed changes will be sent back to Composition, which will make the changes, and the text will undergo a further check. With our current *VIE* set of 44 volumes this is 440 proofing jobs!

To those proofreaders currently with an assignment: could you take a couple of minutes and send me mail (steve.sherman@compaq.com) with a brief summary of your progress so far? The number of Pre-Proofing jobs will be diminishing steadily over the next year, and Post-Proofing jobs may not become available for several months. If you're wondering how to pass the lean times we are entering upon (before the full fury of Post-Proofing is upon us!), I have a suggestion. You will read, in articles that follow below, about the exciting work that is happening in the VIE: Double Digitization and Techno-proofing. Double Digitization is perhaps the single most important effort in the project at the moment. Though we could never have accomplished what we have without the advanced technological tools at our disposal, even these tools are not equal to the task. If we had started two years later, scanning technology would no doubt have progressed beyond its current state and our texts would be cleaner than they are. But if we thought that way we would never have begun the project at all! So we need to use the technology we have now in a way that overcomes its shortcomings, and that is what double digitization is all about. If you have a scanner, or imaging software, or an OCR program, please consider volunteering for this vital effort. It is here that you can make your greatest contribution to the *VIE* at this point in time.

Techno-proofing is an attempt to use technology to supplement the efforts of our wetware proofreaders – who are still our last line of defense. The new tools you will read about have the capacity to isolate words in such a way that incorrect ones can be targeted. My preliminary work with the VDAE has been fruitful, and I have posted the results on the Errata Archive

(you all remember http://www.cs.wisc.edu/~suan/vie/ don't you?).

Finally, let me once again express my thanks, and that of the entire *VIE* management, for the effort and enthusiasm that has characterized the work of the proofreading team. You have all contributed mightily to bringing our excellent enterprise nearer to realization.

Steve Sherman

DOUBLE DIGITIZATION REDUX

Though DD has gotten under way, we do not yet have all the help we want. DD has also proved to be a difficult to understand. It is sometimes regarded as mere re-digitization; *it is not*. DD is not only innovative at each phase, but is in its essences a synergistic process. It combines scanning, image enhancing, OCR and word-processing technology. It is a complex of processes that, taken together, makes up for the impossibility of the VIE to use traditional editing and publishing methods. We are not alone in this "impossibility". Few editors or publishers now use these methods. This is why books and news papers are more and more full of foolish errors and give a growing impression of illiteracy. But DD guarantees that v-texts (our digital versions of Vance's stories in their current state of correction) will suffer from no errors introduced by the VIE itself. Such errors fall mostly into two categories: "scannos" and typos, and confusions and lacunae. Scannos are errors introduced by the scanning and OCR processes. Confusions are mixed-up or added words or phrases while lacunae are 'missing words or phrases'. These errors eliminated, we will only have to correct the typos and editorial meddling already extant in the "preferred" texts, or that portion of it which has survived into the v-texts (a 'v-text' is a VIE text file in its current state of correction). Meanwhile Koen Vyverman and Ian Davis have both designed electronic tools which are helping some of us attack these errors more effectively. Aside from editorial meddling, the most deadly errors of all are wrong words. An example would be: "He want to the dour" for: "He went to the door". Though only proofing can eliminate such errors already present in the published texts, DD can eliminate all such VIE introduced errors. How does DD work? Again; DD is more than scanning. It is a synergistic dynamic between a particular method of scanning, certain kinds of enhancement of the scanned text images, a particular use of OCR programs, and a particular use of Word's "compare" tool.

DD Scanning

DD scanning is not ordinary scanning. The operator must pay careful attention to his equipment, follow the indications we gave in Cosmopolis #9, and work with DD management to squeeze the most out of his hardware, software and procedures. Tests must be made. These include comparing sample images at high magnification, as well as the OCR result of different types of images. OCR results improve notably when the extra time is spent to optimally calibrate the tools.

DD Enhancing

In some cases we will enhance the text images (scans) to produce new images for OCR. This usually means adding contrast or focus to either gray or black and white images (see Cosmopolis #9). Enhancement will not be used in all cases.

DD OCR-ing

The next step of DD is to OCR the DD-quality scans and enhanced scans. Depending on what programs the operators have, or what Richard Chandler can arrange between DD team members, this involves OCR-ing the scan in several OCR programs, or enhancing a scan in various ways and running it though the same OCR program several times. The result is several documents (called "OCR-ings" or 'OCR versions') which not only have few errors in themselves, but, *and above all*, have different errors from each other. This phenomenon, in DD jargon, is referred to as "useful difference". In other words, a given OCR version may have (comparatively) many errors, say 10 per page, while another OCR-ing may only have one error per page but, if this one error is not present in the first OCR-ing, the differences are called "useful". If DD scanning and enhancing is not done to "DD standards" the errors tend to be the same in each version, with the only distinction being quantity. This is the profound importance of enhancement on the one hand, or of using the same scan in different OCR programs on the other. Enhancing takes a file which is read in one way by an OCR program, and modifies it in a way which forces the same OCR program to read it differently. Sometimes this results in more errors but, if the process in done right, the new errors are different from the old. This process is demonstrated in Cosmopolis #9. With a three or fourth DD quality OCR version, all significant errors are wiped out.

Comparing, or Jockeying (DD J)

The "compare" feature of Word makes it a simple matter to compare these texts (OCR versions) against each other and to thus cancel out the errors. By using special screen proofing fonts (like Courier at 14 pts. or VIE devised proofing fonts (RR and Frankenfont)

the operator can also pick up a certain number of errors by eye, though no proofing per se is done in DDJ. The very clean result of the jockeying process is a single text, the "dd-text". Remember: jockeying is not done by the scanners! It is done by DDJ, a sub-team of DD, headed by Chris Corley. In fact scanning may be done by one person, enhancement by another and OCR by another. Richard Chandler, DD team leader, is organizing this work. When the obligatory 3 (or 4) OCR-ings have been generated by whatever combination of scanners, enhancers and OCR-ers Richard can put together (these three might be the same person, or not!) he passes them to Chris, who assigns them to one of his jockeys, who then merges, or 'jockeys', the different OCR-ings into one text: the dd-text. (Chris reports that Rob Gerrand has already completed a jockeying job: the first dd-text!)

Monkeying

The dd-text can then be compared to the v-text. This job, called "monkeying", is reserved to an elite team of VIE text-work veterans, the "monkeys". Again, with Word's compare tool, they study all the discrepancies between the v-text (which has already been human proofed several times) and the dd-text (which is the product of at least 3, DD quality, OCR versions) and sweeps away any remaining VIE introduced errors. We have tested DD for many weeks, and we know it works.

To Resume: DD work has several phases:
1 - Scanning
2 - Enhancing (not always cupatory)
3 - OCR-ing
4 - Comparing OCR versions (jockeying, or DDJ)
5 - Comparing the resultant dd-text to the extant v-text (monkeying)

Monkeying will be done by an elite sub-group of the proofing team under Steve Sherman. Jockeying is done by the DDJ team headed by Chris Corley cjc@vignette.com. Scanning, enhancing, and OCR-ing, will be done by members of the DD team, headed by Richard Chandler chandler@math.ncsu.edu. Join these teams! Members of Richard's team do not need to do both scanning and OCR-ing, which, in the case of some programs, can be laborious. They can simply scan, simply enhance, or simply OCR. It is work that requires learning how to use your particular equipment, whatever it might be, in a new way.

Not all v-texts require DD. Those whose source are the Vances' own digital texts contain no VIE introduced errors; some of the late works are in this category.

Who Can help?

Anyone with a scanner, an OCR program, or imaging software, whatever they may be, can get involved in DD. Even cheap or old equipment, and non-nerd level software, can give DD quality results. It is the operator that counts most in DD. Anyone with a recent version of MS-Word (with the "Compare Documents" feature) can help with DDJ. Consider signing on for a single short story! The innovative nature of the techniques and the high quality of the results makes for interesting and gratifying work.

DD JOB REPORT

I have just DD-ed The Men Return, a 9 page story of about 3000 words. Here is the v-text history:

menre-raw-v1:
scanned by 38 (yours truly), but my scanning method was so poor at the time that it was a 50% typing job.

menre7-raw-v2:
proofed by 408; note that Hans van der Veeke proofed against two versions: TOR88 and NESFA 1985, while I scanned from Ace, and checked against Granada. Hans turned up a number of interesting things, and I turned up some differences between Ace and Granada - mostly in favor of Ace. Tor also seems to have something right where all the others are wrong; but these are issues which TI will resolve.

menre7-raw-v3:
Proofed, again, by 38. But by virtue of the privileges I command as editor-in-chief, after proofing I went ahead (having gotten permission from Richard, Chris and Steve), and did DD scanning (300dpi grey, brightness + 25, contrast +50, gamma + 3) enhancing (created two extra image versions: #1: +100% contrast & conversion to b&w, #2: was #1 "enhanced" twice in a certain imaging program), OCR-ing (TextBridge Pro = 2 versions, TextBridge Classic = 1 version). I then jockeyed (collation of the 3 OCR-ings), and monkeyed (check of v-text against dd-text). Naturally I checked any discovered errors against the preferred text: Ace. The results:

Errors in the dd-text found by the v-text:
1 - single quote for double quote
2 - missing double quote

Errors in the v-text found by the dd-text:
1 - 2 missing commas
2 - 2 missing double quotes
3 - comma for a period
4 - confused word: 'somet imes' for 'sometimes'
5 - inverted words: 'he finally' for 'finally he'
6 - missing letter (insect's')
7 - 'in' for 'into'
8 - missing 'the'
9 - missing 'had'

This is a total of 11 errors, none of them spectacular, but some must be called important. I do not

understand the 'somet imes' error. This should have been picked up by spell check. Perhaps the space was introduced by a false key stroke after I ran spell check, or some glitch when I globally changed the text to Courier to help proofing and monkeying. I have also seen "somet imes" in other documents I have produced, so perhaps this is a glitch in my machine. The 'he finally' and 'in' errors would have been very difficult to pick up by eye, since they read perfectly well. The missing 'had' and 'the' . . . well, neither Hans nor I saw them, but their absence did affect the phrases adversely, as did 'insect'. The missing commas were not too serious, and the comma for the period, coming at the end of a quotation followed by: 'So-and-so said', or some such, was practically nuncupatory. But the missing double quotes look quite poorly if you notice them, as does the single quote for a double quote.

In sum, DD is an absolute must, but our texts are in fairly good order. There were 3 wrong words and two mix ups, or 5 'serious' errors. In a 9 page story this is less than one per page, utterly unacceptable of course! If you count all 11 errors there was more than 1 per page. Let us press forward with DD. DD scanners, in particular, must take special pains!

Paul Rhoads

TECHNO PROOFING

The VIE now has at its disposition not one, but 2 electronic analysis tools to assist text correction: WordPick and the VDAE. We are eager to increase the number of people who use these tools. Following this introduction there are articles about them and their use. For those interested in participating in Techno-proofing here is some background.

Those of us who initiated the VIE did so not on the basis of any profound understanding of editing problems and techniques, but out of enthusiasm for Vance's work. We have had much to learn. But VIE work is also, in many ways, completely innovative, so even the literary pros who have joined us have also had much to learn, or even discover. Among these learned and discovered things is the difference between various kinds of "correction". It took us months to figure out such things as that we had to start our work on the basis of a preferred text, to establish criteria by which to choose those texts, to see that a preferred text, all by itself, will have two kinds of errors (typos and editorial meddling) and to understand fully the difference between these two kinds of errors and how we should approach them.

Typos, once identified as such, can mostly be corrected by the VIE without outside help (meaning manuscripts or input from the Vances). But editorial meddling requires, in most cases, some authoritative source of information, like manuscripts, or the author's say-so. From the latter source we know that "The Dying Earth" is not a Vance title. But when, in the middle of chapter X, of book Y, published 30 or 40 years ago, an editor has cut out a sentence, or changed a word such as 'saloon' to 'salon', these are not the sort of thing the Vances' may remember, or even ever have been aware of. We have also learned that, among manuscripts, there are all different sorts. A first draft is not as useful as a setting-copy, and sometimes the only manuscript evidence we have can not be considered authoritative. Also, and for a long time in the old Merscript days, there was, in addition to productive work, ongoing confusion about what it meant to *correct* Vance. Some people were concerned we should correct the author's "errors", meaning where he used "ungrammatical" constructions, or "wrong" words. The arguments in favor of doing this can seem quite convincing — and in some case it even must be done of course — but it has taken us months and months to sort this out, not only among ourselves, but in our own minds. These issues may seem simple: they are not. The basic VIE policy is, of course, that the VIE is not in the business of correcting Jack Vance, but of preserving and popularizing his work. For the VIE the following holds true: Jack Vance is his own dictionary, his own rule of grammar, and his own style manual.

VIE Introduced Errors

The next aspect of this problem has been VIE introduced errors. We realized from the beginning that we would introduce new errors though digitizing, but at that time we were more concerned about errors in the published texts. Now, however, our concern is just as focused on VIE introduced errors. This is not because they exist in catastrophic quantity — they do not — but because, whether or not we correct all the errors of the preferred texts, and whether or not we bring our texts into conformity with the manuscript evidence we have, if we introduce a whole new set of errors, errors which we thus prove we were incapable of eradicating, the VIE books will be a laughing stock and lack the literary impact we hope to give them. One of our primary goals is, therefore, ZERO VIE introduced errors. We have discovered cases where we have left out sentences or introduced typos and "scannos". Of course the "proof against" (preferred) helps find these errors as much as it helps find the errors (mostly typos) in the preferreds themselves. But Various new methods are helping us attack certain kinds of typos and scannos, including typos in the preferred editions that have made it into the v-texts and have, so far, escaped proofers vigilance. For instance; many OCR programs tend to confuse 'm' with 'rn', or 'n' with 'ri'. These things can be tracked down by searching a text for 'rn'. It is a tedious job, but

effective. A whole protocall of such searches was designed by Chris Corley, who is therefor the spiritual father of Techno-Proofing. Such work is complicated by such phenomenon as the popular scanno 'arid' which can replace 'and'. Such an error can not be caught by a spell-checker, because 'arid', though an error in the phrase: 'this arid that' is a word recognized by the spell checker as correct. DD, of course, will eliminate such errors. DD however, at best, brings the v-texts into conformity with the preferred texts. So here we are back at square one. Conventional proofing is in fact the only way to find words that are not picked up by a spell checker. But many words which are correct in themselves are mistakes in contexts. And while our proofing team is making many findings, there is still more to winkle out, and some 130 texts to winkle it out of!

The Proofing Problem

As emphasized above, DD is important for eliminating VIE introduced errors, but it does little (though not nothing!) to help us find typos in the preferred texts. To give an idea of the scope of the problem presented by these typos in the preferreds, remember the famous 'Emphyrio test'. This was a test taken by about 20 prospective TI people a year ago. We all read Emphyrio, searching for certain classes of errors. Alun Hughes collated the results. The average score was between 20 and 30 catches. The 2 strongest people found about 70 errors each. I myself found 17. Everyone, without exception, found some "unique" errors, meaning errors only they found. I found 2 such unique errors which, percentage-wise, was about the average. What we called errors for the purpose of the Emphyrio text do not exactly match what we are now calling errors in VIE proofreading but, broadly speaking, in a book like Emphyrio there are dozens of typos and other errors, and even if 20 people proof the book, about half these errors will be caught only by one person. The problem is therefore redoubtable. One may say that such errors are not very serious since most people don't notice them; but that is no attitude! Even I, the butt of jokes among VIE managers for my poor spelling and ineffective proofing, am scandalized by the sloppiness of most of Vance's published texts. Vance deserves better.

The New Proofing Tools

Months ago, to help us locate strange characters that might be needed by the VIE Composition team, Suan Yong and Koen Vyverman started creating "utilities" for searching and analyzing the VIE archive. These tools have evolved and are now incarnated in Koen's VDAE. This tool produces an Excel spreadsheet that, among other things, analyses a given v-text in the context of the total VIE archive. These Excel files are meant to be explored interactively and can be used in various ways to pin down errors. Meanwhile, Ian Davis, for reasons unrelated to the VIE, modified a dictionary tool called WordPick in interesting ways. It can turn a text into a word list, and filter this list in various ways by comparing it to other lists. Ian has put his tool at the service of the VIE and has been working with John Robinson, who has long been involved in making individual dictionaries as part of his proofing work. Working with John, Ian has added useful features to WordPick. Patrick Dusoulier, working with Koen, has contributed to the perfection of the VDAE. Though these tools are still evolving, they have both already proved their usefulness to proofing as you will read below.

The Techno-Proofing Team

For various reasons it is not convenient to have more than one proofer work on a text at a time. For Techno-Proofing we will side-step this problem. We are establishing a team of operators, under John Robinson, who will work with Ian's and Koen's tools to scrub down our texts. Such jobs will be done simultaneously with regular proofing since the output will not be up-dated v-texts, but a report of findings (these findings will be incorporated in the v-text by the "monkeys" who are responsible for controlling v-texts against dd-texts). The creation of this Techno-Proofing team will allow the VIE to bring more workers to bare on more texts since Techno-proofing can be done simultaneously with regular proofing. It should be mentioned that "regular proofing" is our last line of defense against certain types of errors.

CUSTOM DICTIONARIES

For some time, in what might be called "proto-Techno-proofing", and along with Chris Corley's typical scanno searches, John Robinson has been augmenting his proofing efforts by creating what he calls *custom dictionaries*. In a recent e-mail he wrote: "My dictionary lists are compiled by making a custom dictionary for each story and then running Word's spell check on it. This gives me a list of all words in the story that Word does not recognize. Some of them are real words, like names of towns or cities on Earth, but most of them are "Vance" words. It is best to make a custom dictionary before starting to proof a story. I print it out and look though it for obvious errors such as mis-spelled words, or words with slight variations in spelling which often signal mis-spellings, and anything else that might draw my attention.

I just finished doing Meet Miss Universe; the Word dictionary *did not* pick up a number of items: words spelled in all caps: "SSEET-TREET", all numbers: "92-14-63-55", or names like: "Miss 44R951", which

appears a few paragraphs later as "Miss 44B951". It took me four reads to see this last inconsistency, and I saw it only because the two versions were on facing pages. The name only appears these two times so there is no way to know if the "B" or the "R" is correct; not an earth shaking issue, but one for TI."

Both Koen's and Ian's tools pick up words that my custom dictionaries do not. Ian's program can also filter a text using several different dictionaries, a special capacity we must explore! Volunteers for Techno-Proofing should understand that this is unplowed ground. We will need to provide each other with a lot of feed-back as we learn how to get the most out of these tools."

John Robinson will be heading up the Techno-Proofing team. Please Volunteer! johnange@ix.netcom.com

THE VDAE

Vocabulary/Dictionary Analysis Engine

Abstract

The VDAE is a tool which facilitates detection of certain classes of errors in v-texts. It churns out an MS Excel spreadsheet containing a list of all words in a particular text, and columns describing various of their attributes. By playing around with the filtering capacities of Excel it is possible to isolate errors.

Introduction

I had originally joined the VIE-team as a back-up archivist with a special task to accomplish (see Cosmopolis 6). Progress in getting the job done has been steady but slow, the latter because of a number of interesting and useful spin-offs encountered along the road. The statistics presented in Cosmopolis 6 were more of the "interesting" kind, while the subjects treated here are in the "useful" category. I will first give an outline of different aspects of the VDAE database as it has gradually grown over the past six months, followed by a section on tokenization, a somewhat dry subject, but important since it provides insight into basic questions of vocabulary analysis: what is a word, and what is not? Two subsequent sections describe a number of "local word attributes" (derived within the context of a single VIE-text) and a few "global word attributes" (derived from the entire VIE database). At the end are suggestions regarding VDAE usage.

Database Machinations

The VIE analysis database has grown within the context of a commercial database management and exploitation tool known as the SAS System. I've been using the SAS System professionally for over three years, and having found its capacity for integration with other software's file-formats (notably MS Word and Excel) unparalleled; it was the obvious choice for my VIE-work.

The centerpiece of my VIE database consists of all the v-texts (the most current VIE-text Word documents). As newly proofed files come in, the powerful 4th generation programming language that is built into the SAS System allows me to automate the manipulation of these documents in specified ways to process the files, creating tables in various sections of the database, updating global VIE tables, and doing comparisons.

One of those database sections has been dubbed the Vocabulary / Dictionary Analysis Engine or VDAE. Its purpose is to build and compare dictionary tables for the v-texts archive. Another section is concerned with 3rd party texts, and yielded the results presented in Cosmopolis 6. Yet another section attempts to apply certain text analysis methods and theories to the VIE-texts as can be found in the literature on computerized text analysis, e.g. by generating Zipf-diagrams. In what follows however we shall focus on the output of the VDAE, which takes the form of an enhanced dictionary dumped into an Excel spreadsheet.

What's in a Word?

When building a dictionary for a given text, the basic question that needs answering is: how to identify a word? The process is called 'tokenization'.

Tokenization is rules by which a 'word' is recognized in a string of text. The question is deceptively simple. Suppose we choose the simplest possible approach: a 'word' is any string of characters surrounded by blank spaces. Sounds OK? Not quite. Take the string <<"Too bad.">>. It would give rise to two entries in the dictionary: <<"Too>> and <<bad.">>, none of which are really desirable. We'd rather just consider <<Too>> and <<bad>>. So nhere must be some pre-processing of the text prior to building a dictionary. We must, in fact, translate into a space (or blank) all characters not considered valid parts of a word. They are then blissfully ignored by the tokenizer. At the time of this writing, the following are, by default, being blanked out:

<< ; : ",. ? ! ( ) < > [ ] { } * ^ @ ~ # $ % _ + = >>.

Hence, in the above example, the double quotes and the period are being ignored, and the tokenizer recognizes <<Too>> and <<bad>> as dictionary entries. But notice what is absent from this list of ignored characters! Most notable are the single quote and the hyphen. The reason we consider single quotes to be a

valid part of a word is to avoid rubbish dictionary entries. Take <<won't>>. If the single quote is ignored, the tokenizer would add <<won>> and <<t>> to the dictionary.

On the other hand, leaving single quotes in the picture may occasionally cause pure noise. Take: <<Glawen had glanced through the 'Syntoractic Primer'>>. This would lead to <<'Syntoractic>> and <<Primer'>> being added to the dictionary. So the tokenizer applies a few more rules: when a candidate token starts with a single quote, it will strip it off. Furthermore, if a candidate token starts with a single quote and also ends with a single quote, it will strip both off. In the sample above this would leave the undesirable <<Primer'>>, but since we'd generally like to leave trailing single quotes where they are (this allows checking for inconsistencies in the possessive form) we'll need to live with this occasional freak. The hyphen is also left in place. This allows composite words to appear in the dictionary. In "Blue World" we don't wish to see <<sea-plant>> fall apart as <<sea>> and <<plant>>. No, we just want to add <<sea-plant>> as such to the dictionary.

A final point is that, beside textual content, Word documents contain many invisible characters. Tab-marks, paragraph-breaks, page-breaks. All those are also, by default, being ignored by the tokenizer!

Local Dictionary Attributes

Once the tokenizer has identified the words in a text, the VDAE proceeds to brew a spreadsheet with one row for each of those words, which it then enhances with columns of information that may be useful when filtering for specifics. Here are the columns currently available on VDAE spreadsheets. It will be hard to follow this section without an actual VDAE spreadsheet open in front of one's nose. But this section will serve as a reference for Techno-proofers once they have been assigned a job and have received the corresponding VDAE spreadsheet.

The Columns

WORD: contains an entry for each string that the tokenizer recognized as a separate item while applying all the parsing rules as detailed above. Without doing anything else, careful study of the word-list can already reveal interesting features. In the spreadsheet for "The Absent Minded Professor" (v-text: ABSEN1) I note the word <<'as>>. In the section about tokenization I explained that leading single quotes are being stripped off. Hence, a word like <<'as>> can only show up if it occurs erroneously in the text with two leading single quotes rather than a double quote character!

PARNUM: may be ignored. It is only useful to me as it refers to the paragraph number of one of the occurrences of the current word. It allows me to link back into different sections of the database.

WORD_FREQ: shows how often a word appears in the current text. Note that capitalization is important. E.g. both <<Actually>> and <<actually>> may appear in a given text, leading to two rows in the VDAE-spreadsheet.

OTHERTHANLOWERCASE: if the current word consists of only lowercase letters, i.e. the range a-z, this column contains a zero. If any other character than a-z is present, it contains the position of the first such character. As an example, consider <<sheriff's>> which has a non-a-z character at position 8.

WORDLENGTH: gives the length of the current word in number of characters.

The following five columns only make sense if OTHERTHANLOWERCASE is not zero. They flag some properties of the current word. If not applicable, these columns contain a period which, by the way, is the generic symbol for missing data in most of my database environment.

HASACAP=1: the word contains at least one capital in the A-Z range.

HASAQUOTE=1: the word contains at least one single quote.

HASAHYPHEN=1: the word contains at least one hyphen.

ALLCAPS=1: the word consists uniquely of capitals A-Z.

ALLNUMS=1: the word contains only numerals 0-9.

VCODE: gives the v-text reference. Not really useful, except for the internal workings of my database.

APPEARS_LOWCASE: will have values 0 or 1. It has 1 if a word not only contains at least one uppercase character, but also appears entirely in lowercase in the current text. If a text contains e.g. both <<Absent>> and <<absent>>, The APPEARS_LOWCASE attribute for <<Absent>> will be set to 1.

APPEARS_CONTRACTED: has only 0 and 1 as possible values. It has 1 for those words containing at least one hyphen, that also appear contracted (de-hyphenated) in the same text. Suppose e.g. that a spreadsheet has entries for both <<sea-plant>> and <<seaplant>>. The APPEARS_CONTRACTED attribute for <<sea-plant>> will have the value 1, indicating the presence of <<seaplant>>.

Global Dictionary Attributes

The final four columns in a VDAE-spreadsheet contain data related to the full VIE analysis database:

VIE_TOT_FREQ: indicates how often the current word appears in the entire VIE database.

APPEARS_LOWCASE_IN_VIE: has 1 if a word not only contains at least one uppercase character, but also appears entirely in lowercase somewhere in the VIE. This could be the current text, or it could be a different text. E.g. <<Burnt>> does not appear in lowercase (<<burnt>>) in ABSEN1, but it does appear so in some other text.

APPEARS_CONTRACTED_IN_VIE: similarly to the above, this column will tell you if a word containing at least one hyphen appears in its contracted form somewhere in the VIE.

APPEARS_ONLY_IN_THIS_VCODE: indicates if the current word is unique to the current text or not. <<Dalrymple>> e.g. is not a word you'll run into in any other text than ABSEN1. It is flagged as such in the ABSEN1 spreadsheet.

Usage

Suppose you'd like to try filtering a VDAE-spread-sheet for proper names? First; turn the Excel AutoFilter on. For those unfamiliar with this feature, you'll find it in the menu Data -> Filter -> AutoFilter. The AutoFilter is present in the more recent releases of Excel. It puts a drop-down list at the top of each column, allowing you to see and select what distinct values are present in a given column. Take care though: at least with Excel97, the maximum number of entries in these drop-down lists is 1000. If there are more than a thousand distinct values present in a column, the filter will show only the first thousand!

From the drop-down filter on: APPEARS_ONLY_IN_THIS_VCODE, select 1. Set APPEARS_LOWCASE_IN_VIE to 0. From the OTHERTHANLOWCASE filter select 'custom' and then 'does not equal 0'. From HASACAP select 1. From HASAQUOTE select 0. Et voilà! Not guaranteed to be complete or entirely correct, but this should give you an idea of what can be accomplished rather easily with the AutoFilter by combining filters on several of the word-attribute columns.

Note that a filter doesn't necessarily need to be that complex. In general, selecting: APPEARS_ONLY_IN_THIS_VCODE = 1 will already produce a list worth having a closer look at . . . Use the other columns to refine the filtering and show only the kind of stuff you're looking for. For more practical examples, refer to Patrick Dusoulier's article below. Patrick has been an enthusiastic user and tester of VDAE-spreadsheets from the start. His feedback has lead to a number of improvements, and his use of the VDAE has lead to the discovery of a number of errors in our texts!

Conclusion

As announced, John Robinson has kindly volunteered to head the VIE Techno-proofing effort, which will have both Ian Davies' WordPick application and the VDAE-spreadsheets at its disposal. Contact John if you'd care to give these new Techno-tools a whirl, to see what kind of hitherto un-annotated errors you can manage to unearth through them.

While John will organize the Techno-proofing work-flow, I will be cheerfully providing VDAE-spreadsheets, and will be happy to give technical assistance, with an ear ready for all suggestions that can lead to improvements in the VDAE!

Koen Vyverman

CRUSHING THE LEMON

Practical Use of the Vocabulary & Dictionary Analysis Engine (VDAE)

What do you do when you're thirsty, and you can't draw any more water from the well? When your "sole meunière" is served, and your lemon has been squeezed dry? When you're about to go out for a date with Julia Roberts, and your toothpaste tube is flat as the Netherlands? You become desperate . . . and therefore desperately inventive: you design a new shape of bucket, you build a com-bined lemon scraper-crusher, you throw the tooth-paste tube underneath a passing steamroller . . . You have no guarantee nor certainty that any of those methods will actually work, but at least you've given it a good try. We face a similar situation with proof-ing: we can't make sure, scientifically speaking, that we've flushed out all the errors, but we want to make sure we've tried our best.

Fortunately, we have in our midst Koen, the Laughing Mathematician! After having erected his Amazing Word Enumerator and written his AWEsome article comparing the curves of Jane Austen with those of Leon Tolstoi, Koen could have chosen to rest on his laurels and sip his usual quota of tequila-based cocktails. But this is not Koen's style. He went on adding extraordinary features, the details of which he describes in the article above! Having completed his contraption, Koen sim-ply threw it to us and said: "See what use you can make of this, guys." And he went away for a fort-night to the "nigh uninhabited western coast of Gran Canaria" . . . His very words! Well, I looked at it, and found that it was good. So did Steve Sherman.

Let me demonstrate some of the ways you can use this lemon squeezer.

The Garbled Proper Nouns Searcher (GPNS)

A fairly frequent sort of typo is garbled proper nouns. In some cases, as you will see later, it's not so much a question of typos but of consistency in the published text itself. Going through a list of proper nouns in alphabetical order gives a good chance of spotting such cases. Settings:

OTHER_THAN_LOWER_CASE=1
HAS_A_CAP=1
APPEARS_LOW_CASE_IN_VIE=0
ALLNUMS=0

I didn't select on "HAS_A_QUOTE", because I wanted to see the occurrences of a proper noun with a possessive case too.

I started with Blue World, containing 7683 distinct words. With those filters, I ended up with 174 words beginning with a cap, and not appearing in full lowercase anywhere in VIE, therefore most likely to be proper nouns. I went through the list, and thought I had found one:

Gallager: 14
Gallagher: 1

Unfortunately, this is OK: there is a "Howard Gallagher, a high ranking police official", one of the founders of the Blue World. Rubal Gallager is one of his descendants, probably, the "h" has been dropped at some time. Still, this might have been a genuine typo, and easy to spot with this method! An encouraging sign . . . So I tried it on Marune: 7806 distinct words reduced to 251 by the filters. I went carefully through the list . . . and BINGO! A really juicy lemon here:

Galligade: 1
Galligade's: 1
Galligades: 1

I checked the text, and lo and behold! there was a missing apostrophe for the possessive case: <Galligades Puppets> instead of <Galligade's Puppets> (chap. 13 page 158, of Ballantine 1975). There is another (correct) instance of Galligade's Puppets in chapter 5. This typo had gone unnoticed, although several proofers had gone through the text, including, to their everlasting shame, two Mentors (I'm one of them). Only excuse they have is that they were not yet Mentors at the time. This was great fun, so I tried the GPNS method on a third text: Palace of Love. The initial 9175 entries boiled down to 385 "proper nouns". Again, the lemon gave some juice, although not so tasty:

Kalzibah: 5
Kalzibahans: 1
Kalziban: 1

In fact this is a TI issue. Kalziban and Kalzibahans are not typos, but potential inconsistencies in the text. The major point is: this was not spotted through 4 successive proofings. The GPNS method also highlighted another genuine inconsistency issue (not a typo), but that one had been endnoted already. I just mention it to show how easily this sort of problem can be detected:

Kouhila: 5
Kouliha: 3

(this looks like a football match result . . . Both almost certainly local teams.)

The Rare Occurrence Check (ROC)

The idea is that an unspotted typo in a common noun has a good chance of having occurred only once in the text. If there were several occurrences, then there's a good chance a proofer would have spotted it already.

This is pure Gallic logic, by the way . . . Also, there's a chance that it may have occurred only in this text, but not necessarily (similar scannos will happen in lots of different texts, of course). So the methodology has to be gradual: Settings for stage 1:

OTHER_THAN_LOWER_CASE = 0
WORDFREQ = 1
VIE_TOT_FREQUENCY = 1

This gives a list of common nouns occurring only once in the whole VIE. Again, I tried Blue World first: I had 112 words out of 7683. It may look tremendously boring to have to go through a list like that, but I find it extremely exciting! The reason is that you find words that never attracted your attention in context, but when isolated, they cry for a look in the dictionary, and it's a marvellous way to expand one's vocabulary. Take "atlatl", for instance. I didn't know what it meant, and I had made a completely naive and erroneous reading of the sentence where it appears in Blue World: " A dart thrower, on the order of an atlatl, was tested, but accuracy was so poor that it was discarded." (Believe me or not, but although I've read Blue World many times, I always took this to mean, literally, that an "atlatl", some sort of pseudo-military rank, had given the order to test the dart thrower! Now I know better . . . I particularly like this word because it's of Nahuatl origin, and those among you who know my nickname will understand why!) I didn't find anything that looked like a typo. Too bad, but that's life ("c'est la VIE", in French!). I then marked all those words in a personal column I had opened, called "check", with the marking "1/1", corresponding to "total Text frequency /Total VIE frequency". I was ready for ROC phase 2:

OTHER_THAN_LOWER_CASE = 0
WORDFREQ = 1
VIE_TOT_FREQUENCY = 2 (that's the progression!)

By a funny coincidence, this gave me a list of 112 words, as in the first setting, but not the same words, obviously. I went through the list with care (I must insist: this requires an effort of will, one tends too easily to skip through such a list and be done with it!) . . . and BINGO! I found the word "tam". Now I would have expected to see "tamtam" maybe, a variant of "tomtom", or "tam" with a frequency of two if there had been "tam tam" (which would have been a typo) or "tam-tam", possibly acceptable . . . I checked in the text: < *He took the mallet, prodded each of the loops in tam, and in turn each of the vanes jerked*.> What a splendid dollop of toothpaste out of the tube! A good scanno that went unspotted by 4 successive proofers (myself included again; another Mentor missed it too . . . I won't tell his name, I'll just say he's had a recent promotion in VIE, but this was before we realised he left typos behind!)

One may wonder why this scanno was not spotted with a normal Word dictionary check: the reason is simply that "tam" is in the Word dictionary! I checked in the Merriam-Webster on-line, and it gave "tam" as in the expression "tam-o-shanter: round woollen cap fitting closely round brows, but large and full above".

After this intense moment of emotion, I marked those words with "1/2" (TEXT frequency/VIE frequency) and tried a new list with VIE Frequency = 3. I found no obvious typo in this list of 107 words, but learnt a lot: did you know that Jack uses the word "pangolay" three times across the VIE? And that a "scalawag" (or "scallawag", or "scallywag" as MSWord prefers) is an undersized, or ill-fed, animal? Bet some of you didn't! At this stage, I had words marked with "1/1", "1/2" and "1/3". The more you increase the VIE frequency, the less likely you are to find a typo, so I stopped there and moved to another phase of ROC: looking at the full list of words, and checking your finds for similar words in the text. It's simple, all you have to do is to find your word, and look at the words above and below. There is a chance you will spot something fishy that you missed when looking at the "rare" word alone. Moreover, it gives you an opportunity to browse through the full list; something may draw your eye. To do that, the setting is simply:

OTHER_THAN_LOWER_CASE = 0.

For Blue World, you get 6740 words . . . but you only have to look at about 330 occurrences. This ROC phase would most probably detect the instances where, for a given word, Jack has used a British spelling once and the American spelling several times.

Not that we care but still, it's worth knowing. It's also a great help to spot inconsistencies in the use of hyphens (now that Koen has solved hyphen conservation), or alternate spellings (scalawag might have been such a case, after all!) Lots of opportunities for "bingoes".

I tried this ROC method on Marune, and got nothing out of it except an expanded vocabulary. I also tried it on "Palace of Love". It gives a funny feeling when, moving down the list of "1/1", you come upon "disgurgled" and "disturgled", but — Shades of Tim R. Mortiss! — then it gave things to chew on: <cafes> in "1/2" mode. Not a typo as such, corresponds to the DAW edition, but it is an incorrect spelling. A TI issue has been raised.

<fete>, a "1/3", already endnoted. But the interest of this method is that we now know that this occurs also in another text: Koen can tell us what text it is.

<objets> a "1/3", and immediately after: <objects>, and thought I had another bingo . . . Not so, it came from <objets d'art>. Disappointing ! Betrayed by a French expression; me!

<pavillions> a 1/1, immediately below <pavilions> (frequency 1, but 28 occurrences in VIE): this one had already been spotted by a proofer (Suan himself!) Tough luck, but it shows the method works.

What A Funny Accent (WAFA)

Another potential lemon squeezer: sometimes the accents on foreign words are not quite correct. Here's a simple way to check accented words, except those where the first letter is accented (but then, you catch those in the GPNS). Settings:

OTHER_THAN_LOWER_CASE > 1
HAS_A_QUOTE = 0

Blue World is a special case: the 56 words conforming to the first criteria have quotes. I checked that none of them had an accented character.

Marune gives a list of 4 words only, one of them being "melée". This may have been spotted by a proofer (I didn't see it when I proofed Marune) but not considered an issue. It's not a typo, it corresponds to the published edition, but I raised it as a TI point. Either there should also be a circumflex accent on the first "e" (orthodoxy) or no accent at all (Anglo-Saxonism!). Suan observed that it is a fairly frequent occurrence in English, the first accent not being thought useful at all. We'll see what comes of it. Palace Of Love gives 9 entries, all correct. You can't win every time.

Conclusion (?)

I hope you've managed to reach this last paragraph. I tried to show how exciting the VDAE can be, and at the same time how it can be effectively used and the

really good results it produces. Still, it may have been a bit hard to follow without an actual spreadsheet in front of you, and the text itself. There's nothing like doing it yourself to get a good grip on it. If you feel you want to know more, I urge you to volunteer for the Techno-Proofing team: you will be in the driver's seat, and I'm sure you will find new and better tricks to squeeze the lemon and crush the toothpaste, not to mention the drops of water raised from the well!

Patrick Dusoulier

WORDPICK

For Those Who Like Lists

WordPick is a computer program that came into existence in the days when computers were unfriendly and difficult to use. There are those who will claim computers are still unfriendly and difficult to use and no doubt they have a point. But I am referring to those not-so-distant days when if you wanted a computer to do anything at all, you had to write the instructions for it yourself. Recently I dusted off my old, original WordPick program and in a spirit of nostalgia began to rework it for today's computer environment.

As a quick overview, WordPick is a list-maker. It reads a text file and from it, makes a complete list of words that the file contains, sorted in alphabetical order. WordPick can make a list from a single file or the list can be the end result of automatically scanning hundreds of files. Once the list is made, WordPick can compare it in a variety of ways with other lists. Some of these comparisons may be to find whether there are any words in List 1 that aren't in List 2.

Another feature is WordPick's ability to "look up" each word in a newly created list and compare it with words in say, a standard dictionary. If it doesn't find the word in the standard dictionary the suspect word is written to a new "special" list. This is a quick way to proofread a document by highlighting words which are either typos or for some other reason, didn't make the standard dictionary.

That, in essence, is what WordPick does. For those interested in the nuts and bolts, more details are given below. But before getting involved in them, there are some questions of a more general nature which are interesting to look at. One of the first is: "what is a word"? Is it just an arbitrary collection of characters delineated by spaces at the beginning and end, or is it something more? I think it is something more. And here we come to the fundamental question: "What are you making the list for?" Once you define its purpose, it becomes easier to define what should be a word and what should not. For the proofreader, any collection of characters is of interest. Perhaps you have text which has been computer-scanned and the software got confused by something and included a word like "thi^%". Proofreaders need to have this highlighted, so in this context "thi^%" is a word of real interest. For the linguist or researcher or maybe just for the rest of us, "thi^%" is of no interest at all. WordPick can home in on these aberrations, excluding them from consideration and including only those which have a valid format. WordPick can then go on to look up this shortened list and further exclude any word which does not appear in a nominated dictionary. For those who want the technical details, here they are:

WordPick is a standalone application written for the Windows 95/98 operating system. It is currently under development so in the tradition of such ventures, inconsistencies can be expected.

WordPick has two main functions:
1. List making (also called dictionary making)
2. List comparisons
List making includes the ability to specify:
- Whether words are "valid" or "include any combination of characters"
- Whether the list includes a frequency count (alongside each word, the number of times it appears in the source)
- A minimum word length. The default is 1 character, but lists can be made which include only words of a minimum of X characters where X is any number between 1 and 50.
- Whether the list should include words with capital letters.
- Whether the list should include hyphenated words
- Whether the list should include words with apostrophes inside them.

Once lists have been made they can be analyzed in a number of ways. List Comparisons include the ability to:
- Define a SOURCE and a TARGET (perhaps two versions of the same file)
- New lists can be created from words which are unique to SOURCE
- New lists can be created from words which are common to SOURCE and TARGET
- Words in SOURCE can be "looked-up" in a number of standard dictionary lists. Anything not found in the dictionary is written to a new list.
- Unique lists found from any of the above comparisons can be RE-SORTED as follows:
(a) Case Sensitive: if words in the source list start with capital letters, they will be grouped together at the top of the list.
(b) Case InSensitive: if words in the source list start with a capital letter, they will be located throughout the list near their lower-case equivalent.
(c) Re-Sorted by Word Frequency. (see above "frequency count") If a list contains a frequency count, it

can be (optionally) sorted on count with the least frequently-occurring words appearing first. The above feature list is not a comprehensive one; it gives an overview only. WordPick is an ongoing project with new features being added as the need arises. If you need further details, please address questions to: Ian Davies, delta1@ihug.co.nz

Ian Davies

Addendum: In a recent conversation with John Robinson, Ian Davis suggested the following terms:
RAW DIC: a list which includes a whole text.
LOOKUP.DIC: a list of all words in the RAW DIC which are not foundf in a standard dictionary
VANCE.DIC: a list also filtered by a list of correct Vance words (John's custom dics).
"If the WordPick RAW DIC of a story is then filtered by both LOOKUP.DIC and VANCE.DIC the resultant list should be quite short and whatever it does or does not turn up will tell us something."

NEWS FROM THE IVORY TOWER

Textual Integrity on the March!

December sees an important milestone in the progress of the VIE, as the Textual Integrity Group moves up a gear with its first TI Conference in Chinon, France (this was going to be called "EuroTiCon", but the ever-alert Patrick Dusoulier warns me this has adverse connotations in French).

The need for textual integrity work — essentially, restoring the texts to the way Jack intended them — has been envisioned from the outset but progress so far has been somewhat limited, for very good reasons.

Firstly, we have had to establish and evaluate the kinds of evidence available to us, an iterative process but one which is now virtually complete. Alun Hughes' articles in Cosmopolis 2, 5 and 7 provide an excellent overview of the kinds of issues we have considered.

Secondly, we have chosen to wait until the bulk of the digitisation has been completed and preliminary proofing carried out before moving on to TI. Our experience has proved that working with texts which are not substantially free from scanning errors is frustrating and inefficient. Now that all texts have been digitised, most proofed several times and with double-digitisation underway to ensure that even cleaner texts can be produced, work on TI can begin to move forward apace.

We have also needed to take time to assess who is best placed to undertake the TI work. While the VIE as a whole is able to use all volunteers, certain additional characteristics are required of TI workers, and the pre-proofing round has been used as an informal shortlisting process. We have therefore been able to recruit the first group of TI workers, many of whom like Steve Sherman and Dave Kennedy have selected themselves by unceasing diligence and a spirit of enquiry. TI is not — and never will be — a closed shop. We welcome expressions of interest from anyone who feels they have a contribution to make. The qualities of the good TI worker include (but are not limited to):

- a track-record of proof-reading excellence
- an enquiring mind
- a conservative approach to textual changes
- an ability to work in the absence of hard and fast rules
- a willingness to commit significant time to TI work

Is TI then an elitist programme? Emphatically so! Inevitably, not everyone will have the combination of qualities necessary for this work. Unlike proofing, TI will be done once and once only on each volume; there will be no safety net. It will require an approach both intellectually robust and artistically sensitive. However, a detailed prior knowledge of textual scholarship and bibliographical issues is not required. One of the main aims of the TI Conferences is to explore these questions in the kind of intensive manner not possible with email. Please do contact Alun Hughes or me, if you'd like more details about what TI work entails.

While signing up for TI is a serious commitment, our aim is attainable excellence rather than impossible perfection.

The first TI Conference will be hosted by Paul Rhoads in Chinon, France over the weekend of 9 and 10 December. Alun Hughes, as TI Group Leader, will of course be present, as will Steve Sherman, Patrick Dusoulier, Koen Vyverman, Linnéa Anglemark, and others. The agenda will cover:

1. Essential bibliographical principles
2. Overview of textual editing
3. How Jack's works were written and how that varied over time (illustrated with Vance ms material)
4. Mechanics of establishing the textual stemma and compiling the
5. Preliminary TI report
6. How to establish the status of supporting evidence
7. VIE Rules of evidence and when to make a TI-correction proposition
8. Use of automated VIE tools to support work.

The emphasis of the sessions will be practical where possible. We will be examining the work done to date on Wyst (where we have had access to excellent typescript evidence) and Madouc, which is currently being prepared for TI by Steve Sherman.

The US Conference will be early in the new year, New Year, at a venue to be determined. One option is Boston, home of the Mugar Library where so many of Jack's manuscripts are stored. There would also be sense in a central location, especially if somebody local

is able to host or otherwise help with organisation. If you are interested in, and have aptitude for, such work, and if you would therefore like to join us at either the European or American conference, please contact Paul Rhoads so that we can discuss it with you. If you are interested in participating in a logistical capacity (moving people, helping with catering) or some other productive way, this would also be welcome. At the very least you'll make new friends from among the ranks of those who, like you, share your admiration for the extraordinary work of Jack Vance.

Tim Stretton: Secretary: Textual Integrity Group

FROM THE TI LIBRARY

Joel Hedlund reports: "Sorry to say; it'll be a while before I can produce any more XIF/TIFF files. An upgrade to Windows ME has knocked TextBridge 9 and Pagis for a loop. Scansoft apologizes and assures us that a fix will be out by the end of November. You may want to pass this enthralling fact on to the other TextBridge users. Another triumph for Microsoft!

Joel Hedlund: Guardian and Stalwart of the TI Library

TYPOGRAPHICAL NOTES

From Gan Starling:

Kindly allow a moment's rant in favor of an all-but-lost aspect of the typesetter's art: the liberal use of non-breaking glyphs.

Non-breaking glyphs lend an air of cast-metal-type professionalism. For instance: the non-breaking space, which never wraps across lines, is used to keep tightly linked words together. In olden days, no decent typographer would divorce abbreviated titles (Mr., Dr., etc.) from their surnames. Seldom would he segregate halves of an uncommon two-word place name. Hyphenated compounds were often treated similarly: kept together on the first few instances. Reading fluidity was thereby enhanced.

Mediocre typography sadly pervades our computer age. Americans, who have ever written telephone numbers with a dash (even verbalizing thus: "One, DASH, eight hundred, DASH, five-five-five, DASH, one-two, one-two") when on a computer substitute periods instead to hold it together on a line. Why? Only because early versions of MS Word disallowed most non-ASCII glyphs. For some of the rest we may blame the pulps, who's publishers were avid for any shortcut whatever to cram more words onto every page of cheap paper.

But this is not worthy of our glorious project! True, people by now are accustomed to seeing lines break where they never did before. Still and all, we have a chance to reassert an all-but-forgotten touch of class by paying attention to such small details. Jack's prose is certainly worth it . . . Don't you think?

From Koen Vyverman:

I've had a bit of a Cosmopolis reading backlog accumulating over the past months, so as I was going through the more recent massive issues I couldn't help but notice the seemingly incessant stream of complaints about the Amiante font.

Frankly, I fail to see what all the hoopla is about. I printed the sample pages from Trullion and inspected them from various distances. I could not discover fault or clashing aesthetics. I read the sample. I found the experience to be pleasant enough.

Asking my wife to read the same sample corroborated my own opinion. She did however find the baseline to be a bit wobbly at times, whereas I had not noticed such a thing. Poring over this together for a while, we concluded that it is reading speed which makes a difference.

I naturally read quite slowly. I like to savour prose and its individual words. My wife tends to read faster, she prefers to get to the meaning of a sentence as soon as possible and get on to the next one. When she did make a conscious effort to read the sample slower, the wobbly effect disappeared.

In view of this finding I would be tempted to suggest that the Amiante font is ideal for its purpose since it almost forces the reader to slow down and pay attention to . . . the words! I have no typographic education whatsoever, so this is of course simply a reader's layman view.

From the Composition Team:

In Milan we were able to see Amiante printed by the machine that will print the VIE, on the paper it will be printed on. This has made possible an in-depth critique, and adjustment of the font is underway.

MATTERS OF SECURITY

Allow me to discuss copyrights, property rights, and how the VIE views its responsibilities to the intellectual property of Jack Vance.

I know that the VIE volunteers and management are committed to treating Jack's property with care and discretion. I know that, for many of you, the comments which follow are "preaching to the choir." But for the sake of the record, and making the VIE's position clear to everyone, subscribers, volunteers, and third parties, I will recite some facts.

Copyrights and ownership of the intellectual properties which are the texts and stories of the VIE are the sole property of Jack Vance. This should be clear to all. No one within the VIE ranks should have

any concerns about copyrights except to think: the stories are the property, lock stock and barrel, of Jack Vance. Naturally, the VIE aims to be his instrument in the disposition of this intellectual property. The VIE has specific contractual obligations to Jack Vance, and it is our intention to honor both the letter and spirit of those obligations.

Curiously, the files of the VIE *do not* belong to Jack. Though he completely determines how we may use them, both through copyright constraints and our agreements with him, the files, which are the product of VIE volunteer work, are the property of the VIE, which is a registered non-profit corporation. Jack determines how we may use his stories, *but nothing else*, since they belong to "us," the VIE.

Why bring this up? Because at least two managers having privileged access to the VIE archive have been approached by people who stated: "You should give me story XYZ since I have an arrangement with Jack Vance to use it for my purpose." *Nothing of the kind!* If a third party has an arrangement with Jack Vance for use of a story; fine. But Jack Vance cannot arrange for the use of our files: they belong to the VIE. Any third party must treat with VIE corporate officers for the files. No one, including me or any other non-officer, or even a board member, has any right to dispose of any files outside of the authorized VIE work flow.

If you are contacted by someone other than your team lead with a request for a VIE file you hold, have no compunction about contacting me, John Vance (president of the VIE board of directors), or Paul Rhoads (editor-in-chief) about the matter. Remember: no one other than your team lead is authorized to see any VIE file you are working on, and this is the only person to whom it should be sent when your job is finished.

Please take this reminder in the spirit in which it is written: the VIE's concern for protecting the property of Jack Vanceand commitment to promoting his interests. You are entrusted with part of the life work of an author for whom we all have the greatest regard — please regard the VIE files with this same respect.

Bob Lacovara

NOTES FROM THE EDITOR

We are, still unofficially, hoping to deliver the VIE sets in Oct./Nov. 2002. We think this is doable, but there can be no relaxing our effort. We must continue to recruit new volunteers, move new people into the management jobs which keep opening up, and find ever better methods. At the moment, and this has been true since the beginning of the project, we have more volunteers than we can put to work. But we should constantly work to reorganize ourselves so that everyone who wants to work can do so. Volunteers, whatever their job, should not dawdle but get jobs finished in a proper period. There is no exact definition for this but, for most jobs, more than a month or two starts being long. Stay in touch with your team head! Managers must keep alert for ways to get a maximum of jobs to a maximum of people, and make sure that all texts are being attended to.

Many new kinds of jobs are opening up. In addition to workers in DD, DDJ and Techno-proofing, we also need an editor for Cosmopolis. Here is a grand opportunity for someone who wants to do important work for the VIE. Bob Lacovara and Debbie Cohen are too occupied by other management duties, to say nothing of their text work, to keep serving as editors, and I have been obliged to fill in this month. We have also put together a team of Cosmopolis proofers from among the proofers who are currently jobless, freeing up other managers who have been doing this work.

The job of Cosmopolis editor is a key management position. To give the image of literary competence appropriate to the VIE, and to maintain and increase the forward thrust of the project, we need Cosmopolis to be energetically and competently edited. It requires a command not only of language skills, but someone with the relentlessness to hound contributors to get in their reports and articles, a spirit of eager promotion and the ability to throw their weight around. We also need someone (it could be the same person) who can do final layout and out-put the document in PDF. Candidates for these jobs should contact me . Don't be modest: include an argument in favor of your candidacy! What we are looking for: dedication and toughness! These jobs take many hours per month, but they are crucial for the project.

The VIE project is all about how much we appreciate the stories of Jack Vance, our desire to thank him for the exceptional pleasure and profit he has given us, as well as to see his work preserved and popularized. The Integral edition itself is a major and noble undertaking, and much of what is published in Cosmopolis is naturally about that work. But Cosmopolis is more than a technical manual and progress report. It is the voice of the VIE, and we want to see it used as a forum for thoughts about what makes Vance so appealing. Cosmopolis would be particularly happy to publish your ideas on any and all aspects of Vance's work, no matter how briefly stated. What is your favorite vancian word, phrase, joke, character, sub-plot, book, genre? and why? How do you think Vance compares to other authors? What are, in your view, the most important aspects of Vance? What makes his writing good? What makes it moving? What makes it funny? What makes it irresistible? Such reflections, no matter how personal, will surely be of interest and help us all to appreciate Vance even more. And if a bit

of controversy is dusted up, so much the better; it is always fun, and sometimes enlightening, to follow debates on such subjects. Cosmopolis is an opportunity for anyone to express their thoughts on Vance; we hope more of you will avail yourselves of it.

Paul Rhoads

THE COSMOPOLIS LITERARY SUPPLEMENT

Issue #3 is available on the Cosmopolis down-load page of the VIE site. It includes new chapters from: *Tergan*, and *Zael*, and to prove how international the VIE really is, a story in French by Raphael Mesa: *Prince Jauquard*. Lis qui peut!

VIE CONTACTS

The VIE Web Page: www.vanceintegral.com

Paul Rhoads, Editor-in-Chief: prhoads@club-internet.fr

Richard Chandler, DD: chandler@math.ncsu.edu

Christian J. Corley, DDJ: cjc@vignette.com

John Robinson, Techno-Proofing: johnange@ix.netcom.com

Steve Sherman, Proofing: volunteer@vanceintegral.com

John Foley, Composition: johnfoley@lucent.com

Alun Hughes, Textual Integrity: a.hughes@newi.ac.uk

Tim Stretton, Textual Integrity: tim.stretton@bigfoot.com

Acknowledgments: The unsigned articles in this issue of Cosmopolis were prepaired by several members of management. This issue was proofed and corrected by Matt Picone, Ron Chernich and Mark Shoulder.

The Fine Print

Letters to the Editor:
Letters to Cosmopolis may be published in whole or in part, with or without attribution, at the discretion of Cosmopolis. Send your e-mail to The Editor, with indication that you'd like your comments published.

Deadlines for Publication:
Deadlines for any particular issue for VIE-related articles are the 21st of the month

Cosmopolis Delivery Options:
- Those who do not wish to recieve Cosmopolis as an e-mail attachment may request "notification" only.
- An HTML version is avalable on the web site.
- The PDF version of Cosmopolis, identical to that distributed via e-mail, is also available on the web site.