# COSMOPOLIS

Volume I, Issue 9 September 2000

## Fans in Japan

Recently, I was sent to Kamakura in Japan in connection with my work on a Japanese spacecraft. As I often do on such business trips, I combined business with pleasure and planned to take a few vacation days for sightseeing.

While planning my tour, I recalled that there were a few ".jp" addresses in the *VIE* e-mail list. It occurred to me that it might be interesting to meet with readers of Jack Vance in Japan. I took the liberty of sending e-mail to several total strangers, explaining that I would be in Japan for a period of time, and asking if they might be interested in meeting, and talking about our mutual interest.

As it happened, two people were available during my stay in Japan. Though my schedule could not accommodate one of the parties, it worked in the case of Hideyuki and Momoe Kanazawa.

My business occupied four days, through a Thursday. With my associates from my firm, I traveled on Friday to visit Kyoto for a day. My companions departed on Saturday, leaving me to my own devices. Between Saturday and Tuesday, I visited Nara, Hiroshima, and Miyajima. I then returned to Yokohama, where I had stayed during my business. Yokohama is convenient to Tokyo, and I planned to spend Tuesday and Wednesday in the Tokyo area.

Wednesday night, my final night in Japan, was a treat. I had arranged to meet Hideyuki at a particular stop on the Odakyu train line out of Shinjuku Station. Let me set the stage for you a bit, by telling you something about Tokyo.

Many *Cosmopolis* readers are familiar with large cities. I myself lived and worked for years in the vicinity of New York, a city so grand and magnificent that, as the joke goes, it is named twice: New York, New York. Suffice to say, though, that I am familiar with New York and a few other lesser American cities, such as Chicago, Boston, San Diego and Philadelphia. I've spent time in London, Madrid and Barcelona, and occasionally even get 40 miles north to Houston, Texas. Large cosmopolitan cities don't impress me with size alone.

Tokyo, however, is another story. Tokyo, I assure you, is New York City on steroids. Tokyo is: huge; dense; modern; and *alive*. In short, it's a hoot as we say in Texas.

The Tokyo subway system in slack hours looks like New York's subway at rush hour. However, the system is simple and reasonable, even for a non-Japanese reading American. In every station, there is some sign with enough Roman names for the tourist to locate the appropriate track and direction. Further, fares are simple: if you can't find the sign telling you the fare (it's somewhere: you have to look!) you simply buy the cheapest ticket. An entry gate takes the ticket, and you take your train.

When you get off, you simply present your ticket to a "Fare Adjustment" machine. It reads the ticket's magnetic coding and tells you how much more money to deposit. In conjunction with the readiness of the Japanese to help a confused tourist, navigating the subway in Tokyo isn't very hard. The crowds at rush hour are not for the complete novice, but still, it's not too bad.

In e-mail with Hideyuki, we agreed on a time and place to meet. Have I mentioned that most streets in Tokyo have no names? The address system used in Japan is rather peculiar to my mind: addresses contain the section of the city, the block number, and then a building number. The block numbers and the building numbers are apparently assigned in the order of construction . . . or some other order: who can say? Not this American . . .

Here's the sum total of Hideyuki's directions for me:

> *Take the Odakyu line from Shinjuku. Then, get off at the 11th station Chitose-Funabashi (next to Kyodo). Please be careful to take the local train. The express train doesn't stop at Chitose-Funabashi. There is a ticket gate ahead (opposite to Shinjuku). I will be there at six with any of my Vance books.*

Ha! No problem! The first time I met anyone in Tokyo, I must have been suffering from the lingering effects of

some sort of cerebral accident, since I agreed to meet a Japanese person in *Tokyo Station* itself, at 9:00 AM! When I arrived I found about 100,000 Japanese men of medium height, medium build, and black hair, all in purposeful motion. (One person stood out: a tall black man, who I could hear speaking fluent Japanese to a ticket-taker. This was so unusual that I walked closer to see the fellow, and realized that he was an associate of mine who worked for NASA. My words: "Gerald, it's a small world."

In any event, Hideyuki's offer to carry a Vance book struck me as droll, since I suspected, correctly in the event, that the conspicuous person would be me. My reply to Hideyuki:

> *OK. That sounds fine. I should not have too much trouble. But you don't need a description of me, do you? ;- ) I am an Italian- American, about 180 cm, slightly overweight, with black hair, turning gray. I will have a backpack, and look slightly lost!*

And so it was. Leaving plenty of time, I made my way to Shinjuku Station, with which I was familiar. (Knowing smirks on the part of those readers familiar with Shinjuku and Kabuchi-cho are, I assert, unwarranted. My familiarity with Shinjuku is due to the fact that the Hilton is there. Trust me.) Finding the Odakyu local train, I boarded, and a short while later arrived at the Chitose-Funabashi station.

I had the impression that I had arrived in what was predominantly a residential and small business part of Tokyo. As far as I could see, I was the only Occidental in sight. This didn't disturb me, by the way. Of all the places I have been, Japan is certainly the safest and easiest to get about. Japanese politeness and helpfulness to visitors is not over-stated in guide books.

Right around 6:00 PM, a Japanese person carrying a quarto volume approached, and indeed, here was Hideyuki. A few minutes later, his wife Momoe appeared. After greetings and a quick stop at a market, we set off for their home.

As we left the vicinity of the station, Tokyo actually began to grow relatively quiet. We walked through narrow streets by residences, and shortly arrived at a pleasant apartment, their home.

Hideyuki and Momoe proudly showed me their collection of works of Jack Vance. They have collected all of his works which have been presented in Japanese, and many

in English as well. The photo with this article shows my new friends in front of their Vance collection, holding Galaxy magazines with Vance stories which I had brought along as a gift. I was presented in turn with a volume, whose cover is reproduced here. It is a Japanese translation of one of Jack's works published by Hayakawa Books, Tokyo: I leave it as an exercise to the reader to guess which one. Curiously, even though it is a paperback, it has a dustcover, which is the artwork you see here. It was originally sold for 250 Yen.


Over a wonderful dinner that Momoe prepared, I asked a few questions. I will paraphrase their answers, since at the time I was busy with my chopsticks, and not a notepad.

*How did one become a fan of Jack Vance in Japan? Isn't he hard to find?* One of Jack's short stories was included in translation in a Japanese book of science fiction. Apparently, one taste was enough for Hideyuki, but finding Jack Vance in Japanese was hard. It was easier to obtain Vance in English editions, but getting those was not a matter of running to a bookstore. Of course, that hasn't changed either in Japan or in the English speaking world.

*What do you like about Jack's work?* The mood and the settings of the stories translate well. The exotic locale, which appealed to both Hideyuki and Momoe, is also a reason, they conjecture, that neither Vance nor science fiction is very widespread in Japan. Their feeling is that Japanese readers are not comfortable with the exotic, and prefer their fiction to be set in familiar climes. This hardly describes Jack Vance, of course. Nevertheless, a significant

amount of Vance has been published in Japanese.

(A publishing curiosity: Jack's works, in Japanese, are about 25% longer, by page count, then the corresponding English version when compared to books of similar page size. Hideyuki suggests that the Japanese printing is just not quite as dense as English printing. In some cases, this means that a Vance novel is printed in two or three volumes where it is found in one or two in English.)

*What problems do the translations present?* Several. Jack uses a wide English vocabulary. The quality of the translation, and the nuances implied by Jack's choice of words, are heavily dependent on the translator. Jack also invents words, using English prefixes, roots, suffixes, parts from foreign languages, and syllables which sometimes sound right to his ear. These present a nightmare to translators. (For an interesting discussion of translation troubles, have a look at the introduction to *The Best Japanese Science Fiction Stories*, edited by John L. Apostolou and Martin H. Greenberg, Barricade Books, New York. )

The translation issues, by the way, play into continuing discussions which wander around the Textual Integrity folks. Often we found ourselves involved in long discussions over the style and delicacy of Jack's phrases. We spent hours discussing the proper time to use specific spellings of "color" as opposed to "colour" and related spelling issues. We discussed Jack's use of commas, and his preferences for the placement of quotation marks. These issues, at least, are *nuncupatory* in translation. That is, no Japanese reader will ever be subjected to subtle variations in spelling; Jack's phrases are translated as best as some translator can do, and in general, translation is a formidable barrier to Jack's prose.

Nevertheless, something in Jack's work comes through to the Japanese reader, and very distinctly at that. Perhaps Hideyuki or Momoe would care to amplify this for *Cosmopolis* readers at some point!

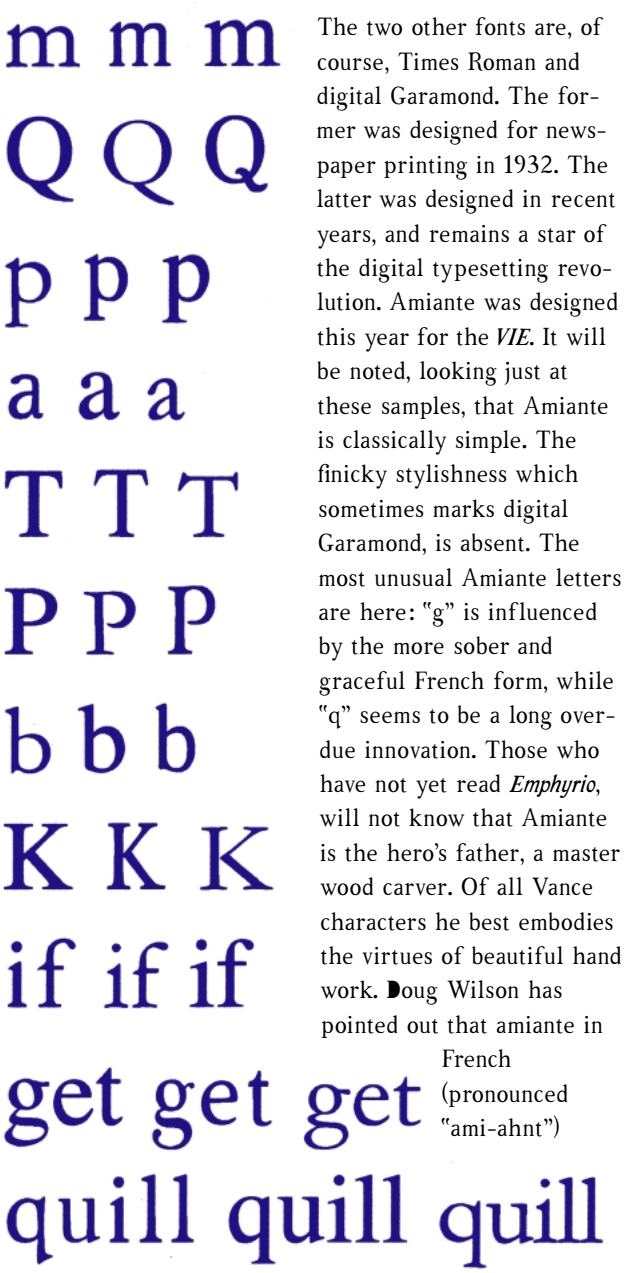
I thoroughly enjoyed my visit with the Kanazawas. We spoke of other matters of interest, and after a pleasant few hours they walked me back to the station. It was easy enough to find my way back to the hotel in Yokohama, with much to think about.

Thursday dawned, and I reluctantly made my way to Narita, and let Northwest fly me home . . .

Bob Lacovara

# Font Test

Test your familiarity with fonts! Can you recognize the letters and words from two popular fonts and the *VIE* font, Amiante?



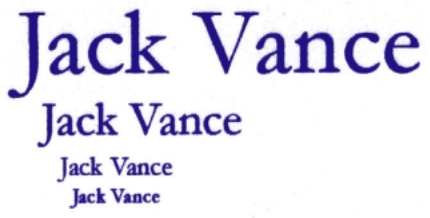

The two other fonts are, of course, Times Roman and digital Garamond. The former was designed for newspaper printing in 1932. The latter was designed in recent years, and remains a star of the digital typesetting revolution. Amiante was designed this year for the *VIE*. It will be noted, looking just at these samples, that Amiante is classically simple. The finicky stylishness which sometimes marks digital Garamond, is absent. The most unusual Amiante letters are here: "g" is influenced by the more sober and graceful French form, while "q" seems to be a long overdue innovation. Those who have not yet read *Emphyrio*, will not know that Amiante is the hero's father, a master wood carver. Of all Vance characters he best embodies the virtues of beautiful hand work. Doug Wilson has pointed out that amiante in French (pronounced "ami-ahnt") means "asbestos" which, in turn, is a Greek word for "incorruptible". We may wonder if this, or the other French word, "ami" (friend) is the origin of this vancian name. In either case, or both, it seems appropriate.

(Answers on next page.)

# Font Sizes

In the preface to *An Introduction to the History of Printing Types* by ◗owding, Alan Burtram writes: "A look at the specimen sheets [offered in the past by typesetting firms] usefully reminds us how subtly or, sometimes, how extensively, the design of metal types was modified in different sizes in order to look the same. Although we have learned to live with it — meanwhile availing ourselves of the many advantages of digital filmsetting — the use of one master for all sizes is one of the more regrettable developments of the last 25 - 30 years. It comes second only to the badly handled adaptation of many PostScript types, a large number perpetrated by producers who should know better."

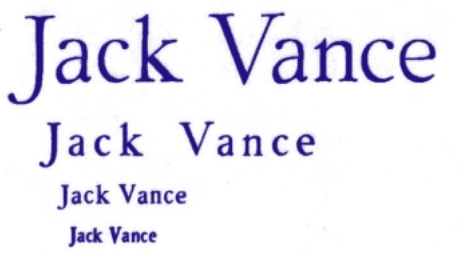
The practice Bartram is referring to gives the following graceless and heavy handed* result as seen in this example of digital Garamond:




By contrast, for each point size needed in the *VIE* format (36, 18, 10 and 8) Amiante has been re-drawn. Here they all are at 36pts.:


* The OpenType Pro format recently introduced by Adobe Systems presents an opportunity to remedy this. These fonts are capable of containing over 65,000 glyphs, and as well as enabling the "optical sizes" mentioned above, are capable of introducing ligatures and other custom elements to the text when used with the proper software.

When used at their proper sizes they appear "the same", or well balanced and appropriate for their size:




(AmianteTitle, 18pt., has deliberately wide spacing, which has been suppressed in the first Amiante example.) Note how the differences between the four versions, so obvious in the first example, disappear in the second. In digital Garamond forms that might be good at 10 points are crude at 18 and 36, while forms appropriate to the larger sizes are reduced to invisible smudges at 8pt.

*Answers to the Font Test:*

| | | |
|---|---|---|
| m – gat | T – tag | if – agt |
| Q – tga | P – tga | get – tag |
| p – gat | b – gta | quill – atg |
| a – atg | K – tag | |

# New DD Phase of the VIE Project

John Schwab and Paul Rhoads have asked me to administer the (new) ◗◗ phase of the *VIE* project.

What is ◗◗? It turns out that our original v-texts are sometimes marred by omitted, altered, added or misplaced words, phrases, sentences, paragraphs, etc., and that human proofing, even with the help of the book itself, is not a sure way of eliminating such errors. ◗igitizing the texts a second time (◗◗ = double digitizing) and then using Word's Compare feature on the original proofed file and the new file has proved to be a valuable method for assuring us that virtually all such omissions/misplacements can be "automatically" caught (and thus corrected). We have further found that with specific procedures and precautions, OCR-ings can

be made to higher standards of quality than our original, raw "v-texts", either scanned or typed. The v-texts have already felt the benefit of human proofing; dd-texts are intended not to bolster the proofing effort per se, but to correct a class of error (mentioned above) introduced by the original digitization process.

DD will require very careful and high quality scanning. If you volunteer for DD you will not be asked to do any proofing. Proofing will not be part of DD. Depending on your equipment, and willingness, you will be asked to both scan and OCR, and perhaps to OCR scans produced by others. I hope all of you are still committed to producing as nearly perfect a *VIE* as is humanly possible, and willing to give more of your time.

The strategy of DD is to produce multiple, different, OCR-ings from a high-quality scan by running it though several OCR programs, or by making a second version of the scan "enhanced" in image programs, or simply by using several scans. We are now in the process of experimenting with the most efficacious methods of scanning and the best use of the various OCR programs. The high-quality multiple OCR-ings we produce will then be "jockeyed" together using Word's compare tool. This will be done by a special DD "jockey team". The errors remaining in each OCR-ing will be automatically cancelled out, while the strengths of each will be combined. The result will be the final

"dd-text.doc". It is this Word file which will be used to correct the v-text. That final task will be performed by a third team.

Remember: DD is not designed to catch typos extant in the texts – that is what proofing is for. DD is not designed to aid TI by helping to compare different editions - that is TI's job and will be done with classic methods. We will, in most cases, be scanning the already scanned text; the "preferred". Our aim is to help deliver correct and compete v-texts to the TI team, so that their work can begin on a solid basis.

The DD program will be fairly intensive. In order to free up TI to do its important work we need to work fast and efficiently, and we hope to complete DD in the next few months. If you are willing to participate, and have scanning capability, the *VIE* will ask you to devote a significant amount of your free time over this period. In order to complete our scanning experiments and efficiently assign tasks, I need to know your hardware/software equipment for digitizing texts. Please provide me with the brand and model of scanner you use and the name and version number of OCR software you have (you can usually find the version number of most Windows software by looking in the "About" field of the pull-down Help menu).

Richard Chandler
Double Digitizing Team Leader
chandler@math.ncsu.edu

# Proofreading Update

The most significant thing to report from the Proofreading Team, at least from my point of view, is the change at the top. Tim Stretton has stepped down in order to become deputy to Alun Hughes in the Textual Integrity team, a post for which he is by education and temperament ideally suited. We proofreaders will miss Tim. He has done a superb job running the *VIE*'s largest team of volunteers and has brought us a good deal of the way towards the completion of Phase I proofreading. For me as proofreader and mentor, it has been a pleasure working with him, and I congratulate Alun Hughes on having won him for Textual Integrity.

Little did I realize that, when I agreed to substitute for Tim during his three-week vacation, I was setting myself up as his replacement. Since I am still finding my way, to some extent, I'm not going to offer the detailed statistics you have become accustomed to reading from Tim. Let me simply say that by the time you read this we should have passed the eight-million-word mark.

Proofreaders: you can help me make this transition as smooth as possible. Those of you with outstanding assignments, if you have not yet responded to my mass mailing, please do so. And I ask all of you to take a look at the page of the website where assignments are sorted by volunteer. (From the main page: Work Documents/Process Integrity/Assignments, by Person, or directly: http://www.cs.wisc.edu/~suan/vie/public/StatusByPerson-c.html). Look at your own assignments. If a task is listed as active that you have in fact completed, please let me know - and if you can, please mail me the file in question. Thanks.

Once I get the active tasks sorted out, it is my intention to make another round of assignments. I thank all of you who have been waiting patiently for more work. I'll be

handing out jobs very soon.

I want to thank Tim for all the help he's given in this transition. By now Tim knows that passing a job onto me means that he gets deluged with emailed questions for a few days, but he bears it with equanimity. Suan Yong has also been of immense help, providing information from his databases. And I must mention Chris Corley and Patrick ◗usoulier, my fellow mentors; and the incalculable moral support of Paul Rhoads.

Last but certainly not least, I want to thank all of you who have written to offer support and encouragement. My favorite remark came from Richard Linton, who wrote: "Needless to say, you've inherited Bureau B, best of luck to you!"

I'm looking forward to working with all of you to complete our excellent undertaking.

Steve Sherman
Proofreading Team Leader

## Scanning 101 and Scanning 102

Scanning 101 is conceived by several professors. Covering the theory, history, and rationale behind ◗ouble ◗igitization, it prepares the student for the technical rigor of Scanning 102 through the establishment of solid ground.

Scanning 102 is presented without prospectus by the estimable and dreaded pedagogues, Paul Rhoads and Richard Chandler. It contains useful research results to be applied in the real world.

These courses are presented in the interest of an informed public, and in the hopes of encouraging participation in the forthcoming ◗◗ effort.

## SCANNING 101
### DOUBLE DIGITIZATION AND OTHER MATTERS

Having spent a whole year digitizing our texts, to say nothing of our multiple proofings, it may seem unnecessary, or even disheartening, to begin the process all over again. But double digitization (◗◗) is both absolutely necessary and not the same thing as our initial (v-text) digitization. The v-texts are our basic texts: with "pre-proofing" (meaning pre-TI proofing), ◗◗, TI, Composition, and "final" (post Composition) proofing, we will hone them into the "complete and corrected works of Jack Vance". It will not hurt to repeat that the word "corrected" must be understood in the *VIE* sense: the *VIE* is not in the business of correcting Jack Vance. We are correcting editorial errors, disfigurations and oversights. The mistakes of Jack Vance, whatever they may be, not including the sort of blunders any conscientious, sensitive and respectful copy-editor would pick up, are not our concern. Little inconsistencies or discrepancies of plot, idiosyncratic use of words, or non-standard grammar and punctuation are not our business, except in that we want to be faithful transmitters of them.

What, then, is ◗◗ for? Traditionally proofreaders work in pairs, reading aloud, sometimes reading backwards. This is the proven way. But the *VIE* is spread all over the world. It is one of our great strengths, but it is a weakness for this aspect of proofreading. We now have enough experience to know that a particular class of error: *missing or altered words and phrases*, has crept into some of our v-texts. Such errors have proved to be beyond the reach of our best workers and even "reads against preferred". How do we know? By pure accident. Enough such errors have been stumbled upon in texts which we thought would be free of them (despite strong volunteers who have done careful jobs) that we have a shadowy outline of the dimensions of the problem. The problem should not be exaggerated, but such errors, even if few, are incompatible with our standards.

Thus ◗◗. What is ◗◗ exactly? It is not redoing what has already been done. ◗◗, for instance, will involve no proofing: proofing is being done to the v-texts, and need not be repeated on the ◗◗-texts. Steve Sherman has pointed out: that the *VIE* project would have been impossible even 10 years ago. Computers, imaging and internet technology are new tools we are using in the service of a passion and a goal. But these tools are still evolving, along with our understanding of how to use them. ◗◗ is about getting our machines to help us as much as they can be made to do; we think ◗◗ can be used to *replace* an aspect of traditional proofreading. We have discovered that, with special care, higher quality

scans can be produced than were produced for the v-texts. Some of this has to do with new software development. We have also made the following discoveries:

- It is always better to scan at 300 dpi.
- It is often best to scan one page at a time, taking the time to square the text and crop close to the text block, eliminating headers, footers, and crease shadow.
- For scanners operated with TWAIN, there are brightness, contrast and gamma settings that can increase a scan's OCR-ability.
- It is also possible to enhance scans in other programs to increase, or simply to modify their OCR capability.

Such modification is important because even if the resultant OCR text has more errors, we have learned that these will tend to be different errors; thus we can apply the ◗◗ principle to ◗◗ itself. By producing several OCR texts using the scanning and enhancing methods mentioned above, as well as by running a given scan through more than one OCR program (the exact combinations will depend upon the optimum use of specific volunteers' equipment) we can produce several usefully different texts or "OCR versions". These can then be sifted together, a process we call "jockeying", using Word's compare tool. Jockeying "automatically", so to speak, cancels out any errors not common to all the versions, while accumulating all that is correct in each.

Since these are not our precious v-texts, the compare feature can be used directly, instead of cautiously throwing the differences up against a reference copy and manually entering changes in the v-text itself, as we will do when the v-text is "diffed" against the dd-text. ◗epending on various circumstances, we plan to jockey together three OCR versions, a process requiring ten to fifteen hours per book, and which will include our standard automated checks as well. To further assist this effort, we are also creating special *VIE* proofing fonts which will make jockeying work even more effective and rapid.

Producing the quality scans required by ◗◗ will take longer than ordinary scanning, but there will be no proofreading - which was the most important time element of regular digitization. We estimate the total *VIE* ◗◗ job at around 1000 hours. We want to keep the ◗◗ teams small to insure quality control, but assuming the scanning and jockey teams have 15 workers each, this is only 33 hours per worker. With 10 workers it is still only 50 hours per worker. It will of course not be possible to distribute this work with mathematical precision, but we believe we have the people and dedication to get ◗◗ finished in a few months, thus freeing TI where much patient work awaits doing. The v-texts will then be compared with the dd-texts by a very small and exclusive team. This will make them "TI ready".

The tests we have run so far are encouraging. With the combined power of the techniques mentioned, and by using the best software, it seems we can get essentially "error free texts". This does not mean ◗◗ will find typos in the digitized book; it will not! This remains a job for proofing. The dd-texts may acceptably include a whole class of minor errors that would not be acceptable in a v-text. What we need them to be is: complete, and without garbled words. We think ◗◗ can be relied upon to sweep the texts clean of *missing or changed words and phrases*. It is an exciting program; we hope to carry it off with élan.

Remember: ◗◗ has nothing to do with TI. ◗◗ is not about comparing different editions. It is about getting a correct digital copy of our preferred edition. Without this basic text, TI can not do good work.

Richard Chandler is heading up the ◗◗ scanning team. He has written to many of you. Please volunteer for this exciting leg of *VIE* work.

# SCANNING 102

## A PRACTICAL STUDY IN DD SCANNING

*Magic is a practical science, or, more properly, a craft, since emphasis is placed primarily on utility, rather than basic understanding.*

*Rhialto the Marvellous*

OCR software is mysterious stuff, and scanning for OCR can seem like using magical "apparatus". We have been doing some serious futzing with our scanning equipment however, and while I, at least, do not pretend to much basic understanding, we are getting some useful results.

There are four, or sometimes five, elements needed to produce an OCR'd text:

- a text to scan,
- a scanner
- an OCR program,
- a user, and
- imaging software.

In the case of each *VIE* volunteer these elements are a unique mix, so for the exigent purposes of ◗◗ it will be impossible to produce a step-by-step instruction manual for the optimized OCR-ing which we need. Each volunteer will have to experiment to optimize results from his own configuration of equipment and software. However, we would like to share some of the things we have learned in the past weeks of intensive experimentation.

My own configuration is nothing outstanding. I have a very ordinary scanner and TextBridge Classic: the baby version of TextBridge that only works with black and white images. My imaging software is a non-nerd, "user friendly" version of Photoshop, and a Font program which has imaging software for black and white images only. I have recently gotten TextBridge Pro 9.0 which has expanded my horizons. This program has some wonderful features, like the highlighting of suspect words in the OCR result, and pop-up comparisons of words with the scan image. Even these great features, however, are not panaceas.

The principal thing we have learned is this: depending upon the character and quality of the image, different OCR software gives different results. *This would appear a notably bland remark, but it is larger than it seems*. While high quality scans are of course a ◗◗ desideratum, ◗◗ also takes advantage of these differences so that the errors from several *usefully different* "OCR-ings" cancel each other out (jockeying). This goal can be approached with three basic techniques, used separately or together:

1– Running a scan, in several altered forms, through the same OCR program.

2– Running a scan though several different OCR programs.

3– Using several scans in one, or several OCR programs.

Which of these techniques will be used in any specific case will depend on who has what equipment, and how it can best be used in the context of ◗◗.

Naturally the best OCR results are obtained from good quality print, on new, white paper, using the latest software. But most of our preferred texts are on crumbling, browned paper (thus the *VIE* project!), and not all the people willing to work on ◗◗ have the latest equipment. But, with careful procedures, even with poor *source material* and *yesterday's technology*, good work can be done.

One of the major problems for OCR turns out to be flaws, imperfections and specks in the paper itself. As an introduction to the problem of scanning for OCR, let's look at such a case in detail. OCR programs can mistake paper imperfections for parts of letters. Here is an example. It is the word '*to*' from a page in my 1981 ◗aw edition of Marune. Note the fleck above the right limb of the '*o*'. This paper flaw is sometimes read as *part* of the '*o*'.

FIGURE 1

My scanner uses TWAIN. This gray-scale image was made at 300 dpi with the TWAIN "Enhance" settings at normal: Brightness=0, Contrast=0, Gamma=1. Note how dark the image is. TextBridge Pro 9.0, which *does* read gray images, does not recognize this image at all. If, however, I make the same scan using settings which I have found effective (Brightness=25, Contrast=50 and Gamma=2.1), TextBridge Pro 9.0 recognizes: 'TO'.

FIGURE 2

To the human eye this image is clearly superior to the first one, and it is to TextBridge Pro 9.0 as well.

Figure 2, modified in imaging software through adjustment of Brightness and Contrast in order to maximize the contrast between word and speck, gives figure 3. This adjustment was made by eye. In figure 2 the contrast between word and speck is already sufficient for TextBridge Pro 9.0 to distinguish speck as speck and word as word. Other OCR software that can read gray images ,however, might prefer the enhanced contrast of figure 3:

FIGURE 3

It is instructive to see that when figure 3 is converted to black and white, the result is:

FIGURE 4

If this image is then "enhanced" in my font program (an operation that fills cracks and eliminates stray bits), the result is:

FIGURE 5

If figure 5 is fed to TextBridge Classic — which only reads B&W — it gives: '*to*'. Success! But if it is fed figure 4, it reads '*td*'. Failure! TextBridge Classic interprets the fleck as the upper end of a '*d*'.

The most important thing I have learned about my scanning equipment, apart from the settings used to optimize OCR-ing, is that scanning in gray is best. This is true even though TextBridge Classic only uses black and white. Even at 300 dpi. and optimum settings, with my configuration it is better to convert an image to B&W from gray, than to scan in B&W to begin with. Doing comparative scans, and inspecting them in high magnification, reveals in detail what the scanner does in its various modes.

For OCR work with TextBridge Classic I have learned that after getting my optimum gray scan — figure 2 — my next step should be to add 100% contrast in my imaging software, before I convert to B&W. If I convert *without* adding contrast the result is:

FIGURE 6

As it happens both TextBridge Classic and Pro 9.0 read this image correctly. However, I know from experience that this class of image does not do well in TextBridge Classic.

When the contrast is added before conversion, the result is:

FIGURE 7

TextBridge Classic reads this image as '*td*'. However, generally speaking, and barring paper flaws, this class of image gets much better results in TextBridge Classic than figure 6, results which are even superior to routine OCR work in better software. Each type of image has is advantages. An OCR-ing of a figure 7 class image may catch 10 errors to 1, jockeyed against an OCR-ing of a figure 6 class image; *but that one error can be important to catch*. In this case the class 6 image would pick up the '*td*' error in the class 7 image. Different image qualities, and weaker OCR programs, can have their DD uses!

I can obtain a further *usefully different* image by enhancing figure 6 with my font program's image tools. Note that, in this program, one "click" of "enhancement" is not enough, and 3 are too many! But two are just right, and the result is this:

FIGURE 8

TextBridge Classic reads this image as '*t6*'. I do not know why this one gives '*t6*' and the previous one give '*td*', but it is interesting to know. Note that these results are constant. If I run the same image though the same OCR program, I get the same results each time. To get different results I must modify the image or run it through other OCR software.

Note the character of the changes wrought in figure 7 by my font program's enhancement operation: cracks are closed and stray bits are eliminated. This sometimes leads to special errors. One example: in the case of double '*l*'s ('*ll*'), if the letters are set too close, or the serifs are smudged, or there is a paper flaw between them, *enhancement* of this sort can join them, and the resulting image can look like a "U" to OCR software. It is important to know what your tools can do, and what is happening down at the micro level. Inspecting the images in high magnification reveals many secrets.

TextBridge Pro 9.0 reads figure 7: as '*to*'. But it reads figure 8 as '*to.*' (note the added period). The more compact dot tricks the software, and this program seems to have been instructed to shift such an out-of-place element, instead of recognizing it as part of a character. The result is that TextBridge Pro 9.0 avoids the '*d*' or '*6*' mistake, but can make a new one of its own: '*to.*'.

Different ways of scanning, or types of post scanning enhancement, improve and degrade OCR results in specific ways depending on the software. With TextBridge Classic the figure 8 class of image generally gives the best results, with errors *usefully different* in character, from those given by figures 7 and 2.

Early in our DD experimentation I was convinced that post-scanning enhancement would be necessary. Recent experiments have suggested that it may be just as

effective to simply use different OCR programs on the same un-enhanced scan. In either case ◗◗ management will want to work with ◗◗ Scanners to help them optimize their work. The above exposé, as far is it directly concerns techniques effective with TextBridge Classic and my own equipment, may be nuncupatory as a model of practical ◗◗ procedure. We may, however, draw from it a few lessons on the level of *basic understanding*:

- The best scan for OCR software is, generally speaking, the one that looks best to the human eye. Letters should be clear, contrast should be sharp, non-letter noise should be attenuated.
- Paper flaws, which are generally faint compared to the darkness of print, are an important source of OCR errors. This fact should be taken into account when adjusting scanner settings: increased contrast is generally helpful.
- In the comparison of images with their OCR result, a minute inspection of the image, as well as the image source (the book page), provides clues to the reasons for errors and can guide the operator to better procedure. OCR software that reads gray images can deal with paper flaws better because the flaws are usually faint by contrast with the print. On the other hand, if the gray image is of poor quality the image might be partly, or even totally, illegible, as is figure 1 in TextBridge Pro 9.0.

Paul Rhoads

# OmniPage (OP) Hints

I use OP 10.0 with a TWAIN scanner (Epson Perfection 1200) and these suggestions apply to that combination. I don't know how much of this will work with earlier versions of OP but it should be worth checking out.

Old yellowed paperbacks (the kind of Vance versions most of us have) pose a real problem since the scan picks up a lot of extraneous garbage from the page. OP Help suggests scanning in Grayscale rather than in Black and White (B&W). This has worked for me but it is dramatically slower on my scanner. I have discovered that the quality of B&W output from OP can be dramatically changed by a few settings, and I would try this before resorting to Grayscale:

- Under "View" select Toolbars and check all three.
- In the "Tools" menu click Options.
- Under "OCR" move the slider all the way to "Most Accurate".
- Under "Scanner" set Page ◗escription "Size" to Letter and "Orientation" to Portrait.
- Under "Process" select "When bringing a new image insert after last page"; select "Automatically straighten page image"; and select "Automatically correct page orientation".

The crucial setting is "Brightness". You will need to experiment with it. For me the most accurate setting is about 2/3 of the way toward the "◗arken" end, but this will vary a little from book to book. The key word here is "experiment". I try scanning the same page at different brightness settings to see which results in the best OCR output. OP remembers the most recent setting.

Once I begin scanning in earnest, I do one chapter at a time. I work in Manual OCR rather than in AutoOCR or OCR Wizard because these last two settings result in too much to clean up later. There are 3 little pull-down menus to the right of the "Manual OCR" button. In the first select "Scan B&W", in the second select "Original Layout and Output Format", and then choose "Mixed Page" and "Remove Formatting". In the third select "Copy to Clipboard". Once these are set, scan the chapter a double page at a time: lay the book flat on the scanner so that a double page is on the scanner glass. Click on "Scan B&W". Repeat this, one double page at a time, until the chapter is finished.

Now comes the fun part: OCR-ing. It is better to use manual zones rather than let OP select the OCR zones because you can eliminate page titles and page numbers (and the shadows, specks, etc., at the page crease) which you don't want anyway. I find it very worthwhile to use a Zone Template. To create one go to the OP Help, search for Zone Template and follow the instructions given under "Creating and using Zone Templates".

I take a typical double page scan and draw a zone around the text on each page, excluding the title line and page numbers. Select the first double page in the Thumbnail View. In the Image View, OP places the zones you have described in the zone template (you may have to drag them around a little for accurate placement). In the "Process" menu select Recognize Current Page, and under "Zoning Instructions" select Use Only Current

Zones. OP OCR's the text and when it is finished, repeat this for the second page, etc. When all pages of the chapter are OCR'd, click "Copy to Clipboard", go to Word and paste. Under the "Edit" menu in Word, select Replace, and click "More" if the Special button is not showing. Under "Special" click "Manual Page Break", then "Replace All" (making sure the "Replace With" field is blank). Repeat this for "Section Break".

◗o the next chapter . . .

Richard Chandler

## New Organization

There has been quite a bit of behind the scenes shake-up this month in *VIE* management. ◗ebbie Cohen has added the editorship of *Cosmopolis* to her gatekeeper duties to free up Bob Lacovara for text work. To boost TI, a job that keeps getting bigger, Tim Stretton has moved from Proofreading to assist Alun Hughes. Steve Sherman is therefore taking over Tim's old duties. Chris Corley and Patrick ◗usoulier will remain proofing mentors, but they also will have special ◗◗ duties: Chris will run the Jockey team, and Patrick, with Bob, will be delegated to diff v-texts against dd-texts. Koen Vyverman will help link the jockeys to the archive, as well as running some special pre-diffing tools he has created that will give Bob and Patrick a leg up on their job, as well as helping us track ◗◗ over all.

## The TI Library

Also to optimize TI work, we are creating the *TI Library*. Joel Hedlund will head up this effort, but he will need help. The task of the TI Library is to supply TI workers with copies of those texts they will need to do their jobs. We hope that *all VIE volunteers and subscribers* will contribute to this effort. Joel has discovered that a book can be scanned in only 2 hours (this for the purposes of the TI library reading copies, not ◗◗ OCR scanning). This scan can be saved in the compact XIF format. The XIF reader, a free download from ScanSoft (http://www.scansoft.com), is an excellent tool which allows a XIF file either to be conveniently read on-screen or printed out. The TI Library will provide books or photocopies if need be. TI will be informing Joel and his helpers which texts are needed, and the TI library, taking advantage of your helpfulness, will get the texts into the hands of the right people. You will all be hearing more from the TI Library.

## Are Subscribers Customers?

The *VIE* is a non-profit business. What is the goal of the *VIE*? It is the promotion of Jack Vance so that the world may benefit from this completely exceptional author, the neglect and incomprehension of whom is one of the great scandals of literary history. We of the *VIE* project may think of ourselves as the stone workers employed by Pharasem the sorcerer to create his "great Talisman". In this metaphor the *VIE* book sets are like the carved forms of Pharasem's "great talisman", and the contemporary literary elite is TOTALITY. Subscribers, like digitizers and proofers, are "volunteers" who are helping make the *VIE* happen. Our project is in fact a gigantic publicity effort to put Jack Vance on the literary map, where we all agree he ought to be. The books themselves will be a highly exclusive but symbolically important result of our effort. We also believe these books will have an influence out of proportion with their number. They will certainly carry the *VIE* goal into the future; their sheer existence, diffused over the entire Terran globe, will help promote Vance's work as long as they exist; and they will be built to last.

But the *VIE* books will be designed to do more than last. The set as a whole, as well as each book in particular, will be designed to speak the *VIE* message. At the heart of this, of course, are our respectfully correct and restored texts, with Jack's own titles. But each physical aspect of the books will also speak the *VIE* message. The books will not say: "Jack Vance is a science fiction writer". They will not say: "Jack Vance is a 20th century writer". They will not even say: "Jack Vance is a respectable writer like many others." No. We are not going to all this effort to make any such tepid announcements! Instead the *VIE* books will cry out: "Jack Vance is a great artist, a classic for all times, places, ages and degrees: he is exceptional!" This message will be implicit in everything from the paper choice, to the

format, to the font. There will, of course, be nothing showy about any of this; quite the contrary! It is through sobriety, simplicity, and true taste that the physical aspect of the books will speak.

The aesthetic considerations which guide us are naturally an object of legitimate interest to all *VIE* volunteers, proofers and subscribers alike. Recognizing this we have steadily provided pertinent information. As soon as we have a mock-up volume from Sfera (later this year perhaps), photographs of it will be posted on the site and published in *Cosmopolis*. The font in particular has generated both excitement and concern. It has been advanced that the goal of the *VIE* is compromised by using an unconventional font, which some even disapprove of on artistic grounds, claming it will impede people from buying or reading the books.

*VIE* management respectfully rejects this hypothesis. It is quite possible that later mass-market editions will make use of the *VIE* texts; we certainly hope they do. In this case, since future mass-market publishers may be prisoner to the focus-group mentality, such editions may well be set in common-place fonts. But there is no good reason for the few hundred books we will print to be thus disfigured and rendered typographically banal.

While we respect the opinions of those who do not like Amiante, we do not share them. We consider this font to be superior to other available fonts for our purposes. The Vances themselves strongly support it. A *VIE* specific font is, of course, a tribute to Vance's work. Amiante, however, was not chosen for this reason but because it has those classical qualities which are appropriate both to Vance's prose and the *VIE* message. Admittedly such things involve a measure of taste. But, of one thing the world may be sure: mass taste, whatever it may be, will not dictate to the *VIE*. Mass taste has done nothing to promote Vance up until now; quite the contrary. The *VIE* will treat and present Vance's work as "great" and "timeless", and the *VIE* project will rely upon its own intelligence and creativity to reach its true goal.

A final remark; the editors of *Cosmopolis* encourage all *Cosmopolis* subscribers to contribute their ideas inspired by the work of Jack Vance. Our goal is not the promotion of any particular interpretation of Vance's work, but to demonstrate to the world that Vance is a rich, thought-engendering artist. In *Cosmopolis* we hope to build a body of thoughtful commentary on Vance's work which can serve as a starting point for future literary studies. These commentaries may even be the object of an actual publication. Paul Rhoads has been the most active contributor of such texts, but we are also proud of the excellent articles by Alexander Feht and Timothy Virkkala, and we invite more of you to share your thoughts on the work of Jack Vance.

The Editors

# Who We Are

The Who We Are pages are un-backlogged. If you wish your bio to be updated, or have a photo to send, or were missed completely, please email pertinant information to: chaschcity@hotmail.com.

# Finally

Here ends the September 2000 issue of *Cosmopolis*. Like Steve Sherman, I am finding my way, and here apologize for aberrations in procedure and/or any inadvertent insults. (I should especially not wish to insult Texans.) At any rate, I hope to be de-aberrated by next month. Thanks to all, especially Bob Lacovara, Paul Rhoads, and Joel Anderson (who is our type-setter) for help/advice/moral support, and to Chris Corley, Bob, and Joel Hedlund for comic relief.

*Cosmopolis* is a group effort. As always, letters, articles, and comments are solicited. You are invited to participate, and always welcome!

Deborah Cohen

Alias Rex

## *VIE* & *Cosmopolis* Staff

The *VIE* Web Page
www.vanceintegral.com

Deborah Cohen, Editor of *Cosmopolis*
chaschcity@hotmail.com

Christian J. Corley, Assistant Editor, *Cosmopolis*
cjc@vignette.com

R. C. Lacovara, Editor Emeritus, *Cosmopolis*
Lacovara@infohwy.com

Paul Rhoads, Editor-in-Chief of the *VIE*
prhoads@club-internet.fr

John Robinson, Publicity Coordinator
johnange@ix.netcom.com

Deborah Cohen, Volunteer Coordinator
volunteer@vanceintegral.com

### Team Leaders:

Richard Chandler, Double Digitization
chandler@math.ncsu.edu

John Foley, Composition
johnfoley@lucent.com

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

John Schwab, Digitization
jschwab@uswest.net

Steve Sherman, Proofing Text Entry
Steve.Sherman@Compaq.com

# 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 Bob Lacovara, 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, but for short story inclusion I must have your copy by the 14th. If you have any questions about publishing your story in Cosmopolis, drop me an e-mail.

### *Cosmopolis* Delivery Options

There are two delivery schemes for *Cosmopolis* readers. Those of you who do not wish to have Cosmopolis arrive as an e-mail attachment may request "notification" only. When a new issue of *Cosmopolis* is ready for distribution, an HTML version is prepared for our web site, and it may be viewed there.

A PDF version of *Cosmopolis*, identical to that distributed via e-mail, is also available at that site.

### Publication Information

For reprint information, address Bob Lacovara.

*Cosmopolis* is assembled, edited and transferred across the Gaean Reach from Minneapolis, MN, and Merrimac, WI, United States of America, Sol III.

*Cosmopolis* is delivered as an Adobe® Acrobat® PDF file. If you wish to have the most current version of the free Acrobat reader, follow this link:


http://www.adobe.com/products/acrobat/readstep.html