

Preparing legacy taxonomic texts for semi-automated mark-up

User Manual
Ver. 1.3

Thomas Hamann

Copyright: Document copyright © Thomas Hamann/Naturalis Biodiversity Center 2013-2016. This document is licensed under a Creative Commons Attribution-ShareAlike 3.0 Unported (CC BY-SA 3.0) license.

This project was subsidized in part by the EU project “pro-iBiosphere” (Grant agreement 312848).

Introduction

Before legacy taxonomic works can be semi-automatically marked up using Perl scripts, they need to be cleaned up and converted to a format that the Perl scripts can handle. This document explains the various steps of this process, including the problems that can occur and how to solve them.

Table of Contents

Preparing legacy taxonomic texts for semi-automated mark-up	1
Introduction	1
Before you start	3
Source documents	3
Required programs	3
Windows set up	4
Microsoft Word set up	5
Notepad++ set up	10
Extracting text from Adobe PageMaker or InDesign files	11
Extracting text from PDFs	14
PDFs with good-quality OCR	14
PDFs lacking a good quality OCR or having no OCR at all	17
Cleaning up a legacy taxonomic text	22
Safeguarding the original file	22
Actual clean-up process	23
Overview	23
What to remove or fix?	24
Addenda, corrigenda, and emendanda	24
Graphics	26
Removing graphics - basics	28
Removing graphics in text boxes	28
Fixing unrecognized text	30
Preparing graphics	32
Unnecessary text	32

Headers and footers	33
Tables of contents.....	35
Indexes	35
Text formatting that interferes with mark-up	37
Interrupted text	37
Missing paragraph marks.....	40
Page, section, and column breaks	40
Text in text boxes.....	44
Footnotes using Word's footnote feature	46
Manual line breaks.....	49
Column issues	50
Tables	50
Lists using Microsoft Word's list function.....	55
Text wrongly recognized as a list	57
What can be left in?	57
Renumbering indented keys	58
Removing styles	61
Processing order within the document	62
Preparing files for combined mass-processing.....	63
Saving the cleaned up text file for processing with Perl scripts	63
Appendix I: Perl scripts to further prepare the text	65
Running the clean-up script.....	65
Running the OCR error fixing script	66
The final manual preparation step: Separating taxa	66
Appendix II: Additional text preparation prior to script running for Flora of the Guianas	69
Initial file clean-up	69
Preparation of "Collections studied"-index for easier mark-up.....	70
Moving "Collections studied"-index and wood descriptions	76

Before you start

Some preparations should be made before you start. These consist of checking whether you have the proper type of source document available for the legacy taxonomic texts, and the set-up of the programs you will use.

Source documents

Prior to preparing the text of a legacy taxonomic work for semi-automated mark-up using Perl script, the taxonomic work should have been scanned in at a resolution of at least 600 DPI. Then Optical Character Recognition (OCR) should have been applied to the scans, with special care taken to ensure that the special characters present in taxonomic works (e.g. male and female symbols) are properly recognized.

This manual assumes that you have the OCR'ed text of the whole legacy taxonomic work available as a Rich Text Format (RTF; **.rtf** extension) or Microsoft Word (**.doc** or **.docx** extension) file.

For more recent legacy taxonomic works that are from the last 20 years and were produced digitally, Adobe PageMaker or InDesign format is also acceptable.

If you only have PDFs, you will need to extract the text from those files by copying it or export the PDFs in a suitable format.

Required programs

The following computer programs are required for the text preparation process:

- Microsoft Word 2010 or later
- Notepad++ (<http://notepad-plus-plus.org/>) - there is a Portable Apps version available in case you cannot install your own programs.
- Latest version of ActivePerl Community Edition (<http://www.activestate.com/activeperl/downloads>) - see **script use.doc** for installation instructions.
- Optional: Adobe InDesign CS2 or later, for legacy taxonomic works that are available as older Adobe PageMaker or InDesign files.
- Optional: Adobe Acrobat Pro or an OCR program (preferred, e.g. ABBYY FineReader Professional) that can handle PDFs with text with special symbols.

It is assumed you are running a version of Microsoft Windows, preferably XP or higher.

If you are not running Windows, and instead are an user of Mac OS X, Linux, or any other UNIX-like operating system, you will need to find replacement programs, and you will be on your own to make everything work. However, hopefully this manual is clear enough to get you started.

(continued next page)

Windows set up

Windows should be set up properly to facilitate your work.

- 1) Go to the “Control Panel”, and check in the “Folder Options” control panel under the “View”-tab (Figure 1) that “Show hidden files, folders, and drives” is checked, and that “Hide extensions for known file types” is unchecked.

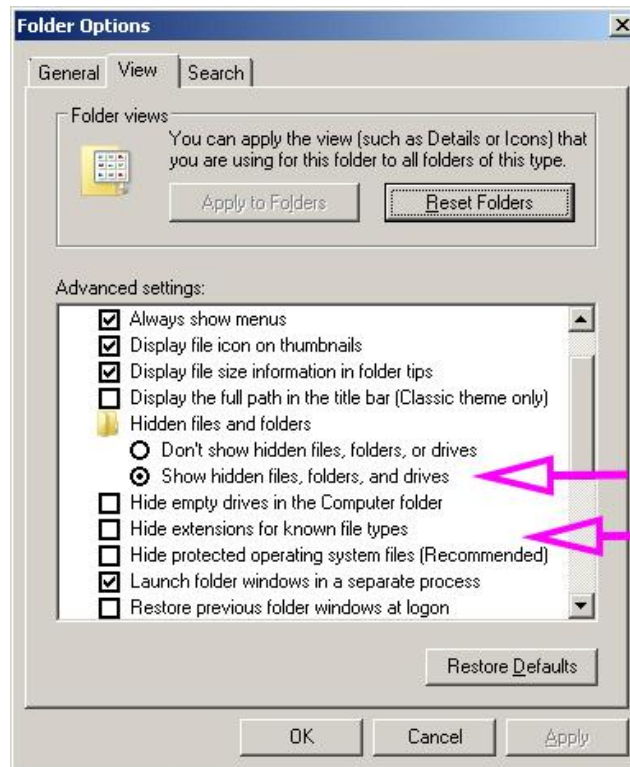


Figure 1: Important options to set in the “Folder Options” control panel.

- 2) In the “Region and Language” control panel, under the “Keyboards and Languages”-tab, choose “Change keyboards” and ensure that you are using a “US” keyboard layout, not the “US-International” keyboard layout (Figure 2). This disables the smart quote and accents feature that makes typing programming languages cumbersome.

(continued next page)

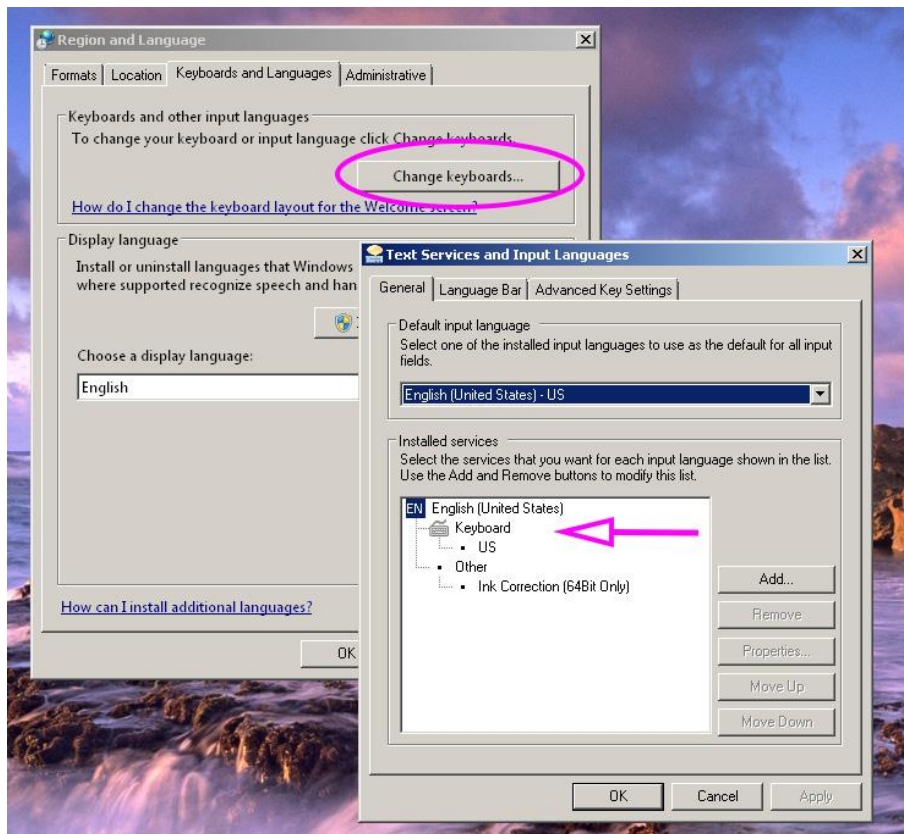


Figure 2: Proper keyboard settings.

Microsoft Word set up

Some options in Microsoft Word should be changed to facilitate your work.

- 1) Click on the "File" tab, and click on "Options" (Figure 3). The Microsoft Word Options window will open (Figure 4).

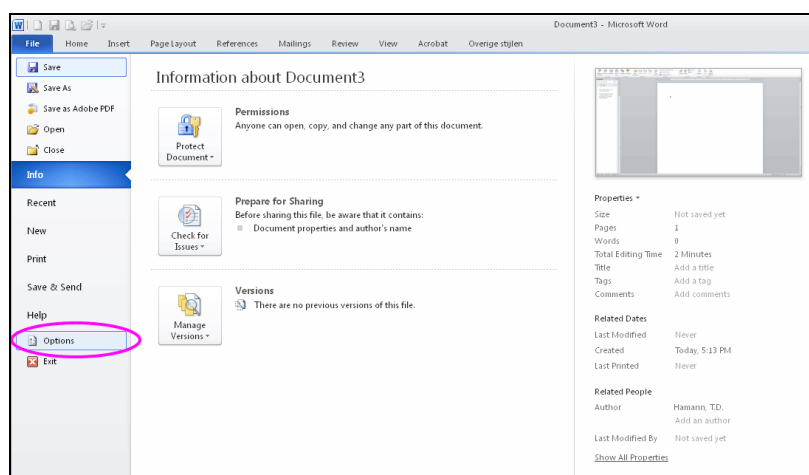


Figure 3: Where to click to open the Word Options window.

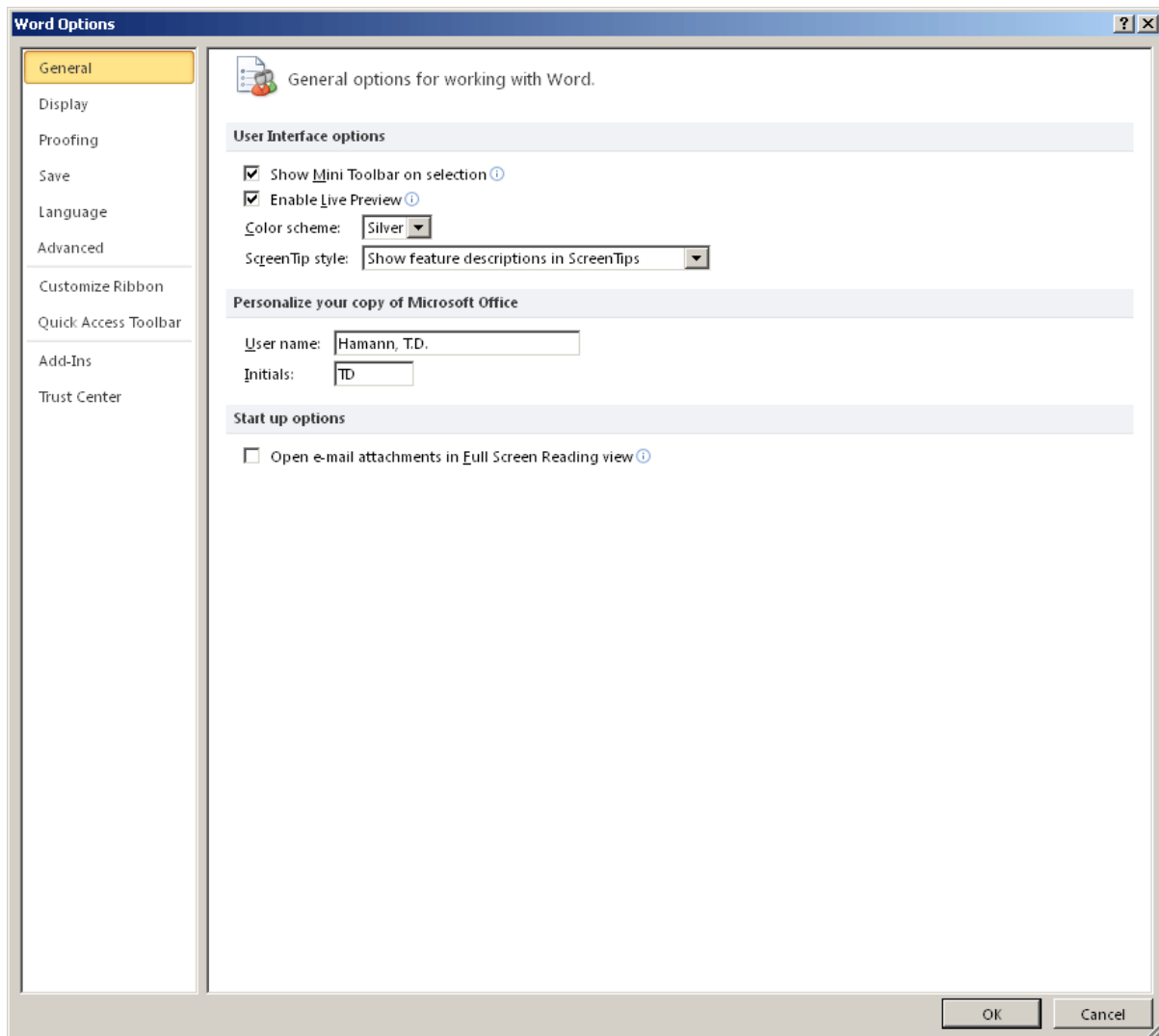


Figure 4: Microsoft Word Options window.

- 2) In this window, click on "Display" at the left side (Figure 5).
 - a. Under "Page display options", look whether "Show white space between pages in Print layout view" is enabled (if not, check it).
 - b. Under "Always show these formatting marks on the screen" enable "Hidden text", "Object anchors", and "Show all formatting marks".
- 3) Optionally: Now click on "Proofing" at the left side of the window (Figure 6).
 - a. Under "When correcting spelling in Microsoft Office programs" enable "Ignore words in UPPERCASE", "Ignore words that contain numbers" and "Ignore internet and file addresses".
 - b. Under "When correcting spelling and grammar in Word" disable "Check spelling as you type" and "Mark grammar errors as you type".

(continued next page)

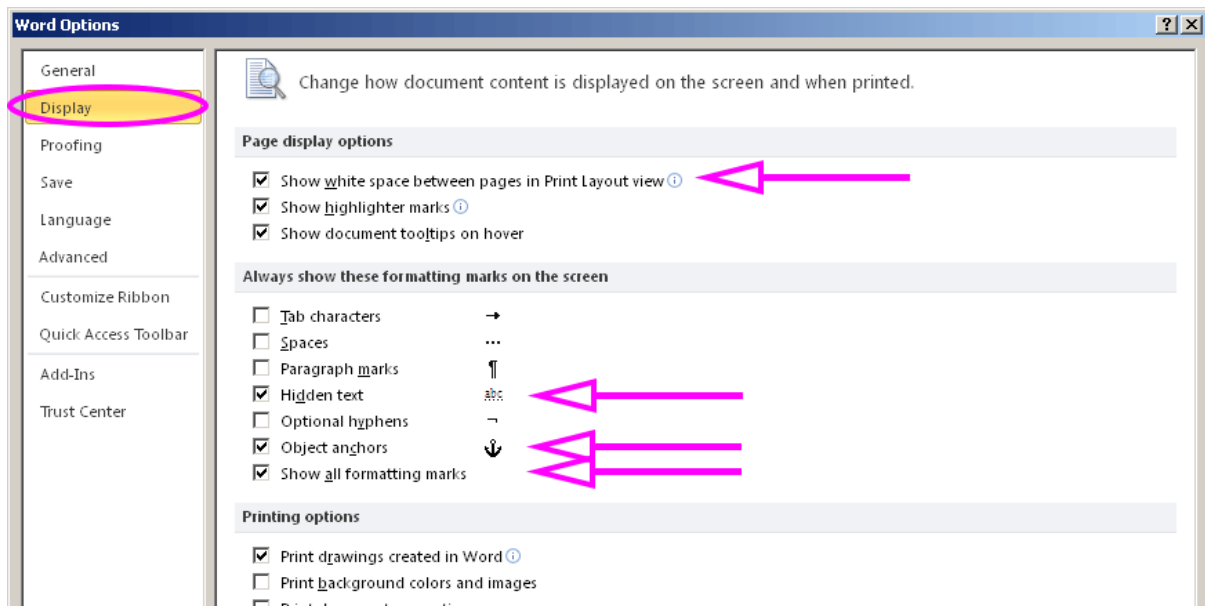


Figure 5: Display options.

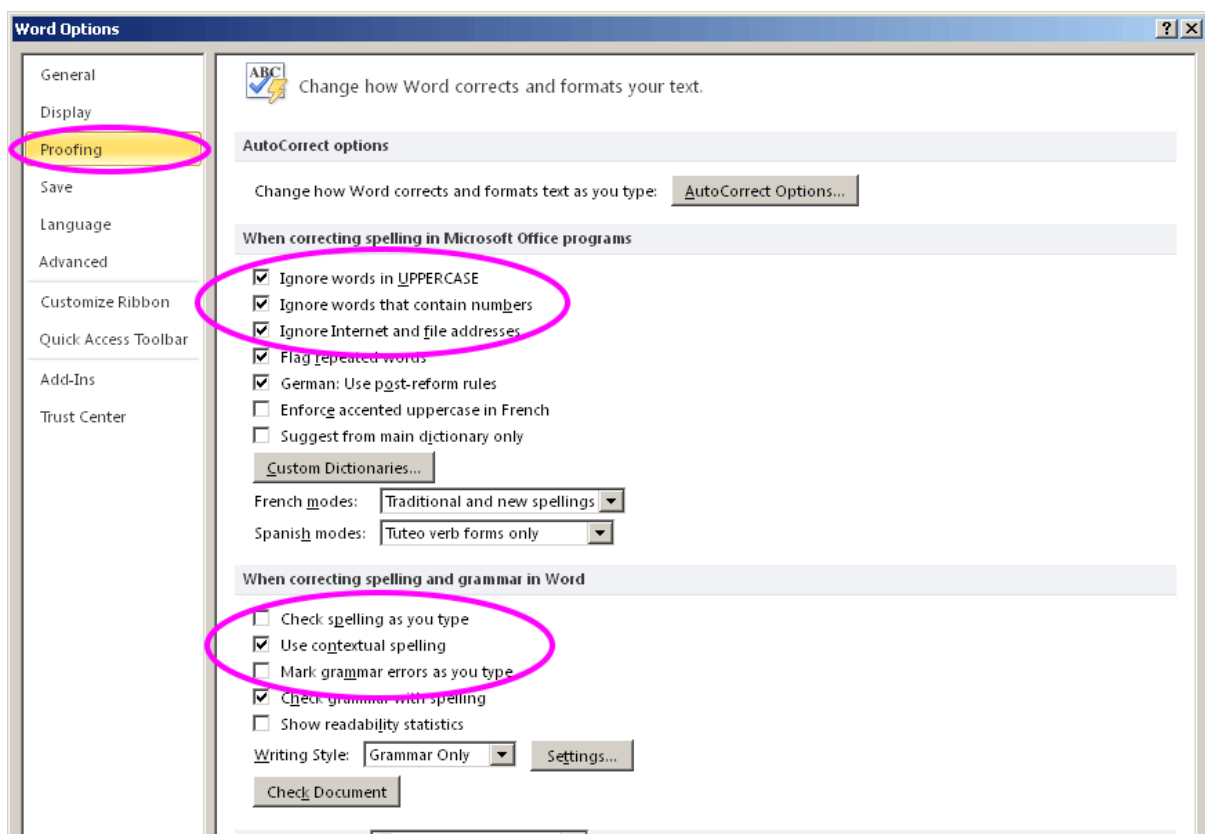


Figure 6: Proofing options.

(continued next page)

- 4) Click on "Save" at the left side, and enable the Autorecovery option (Figure 7).

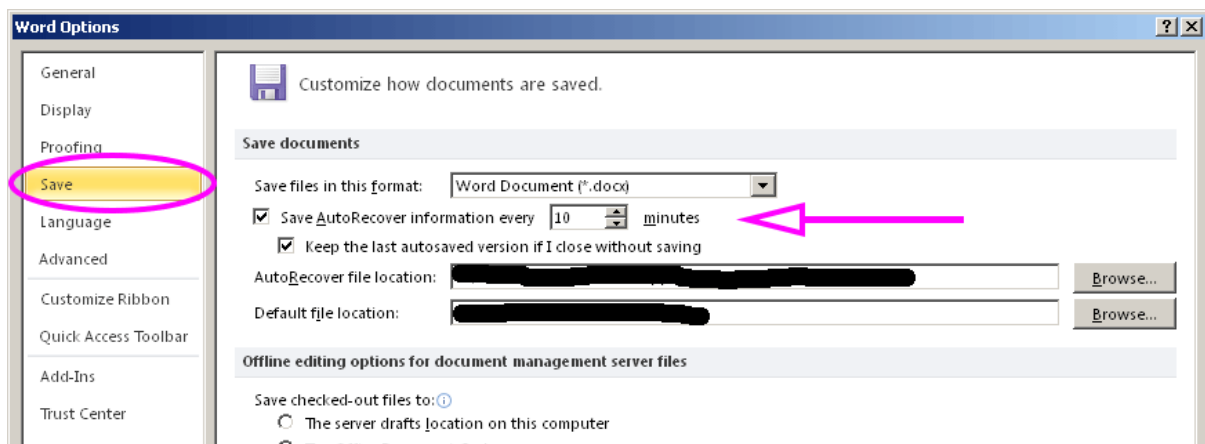


Figure 7: Save options.

- 5) Click on "Advanced" at the left side (Figure 8). Now under the "Editing options":
- Disable "When selecting, automatically select entire word".
 - Enable "Allow text to be dragged and dropped".
 - Disable "Use the Insert key to control overtype mode" and also disable the option right below it.

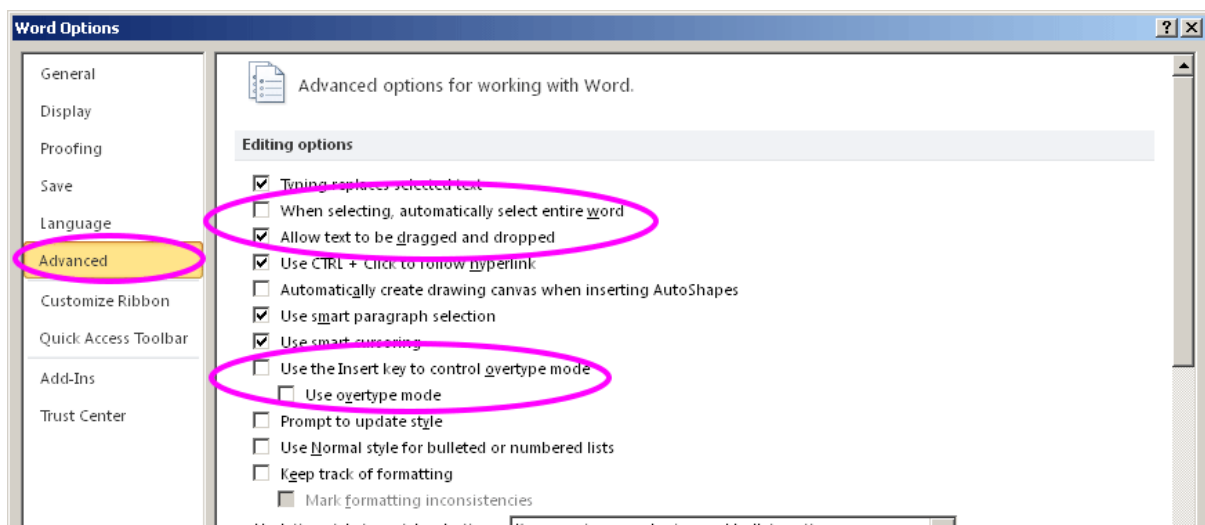


Figure 8: Advanced options.

- 6) Now click on the "OK"-button at the bottom of the window. The "Word options" window closes while saving your new settings.
- 7) Click on the "View" tab and check whether you are in "Print Layout" (Figure 9). It is possible that when you open a document in Microsoft Word the view gets changed back to something else. In that case, do this step again.

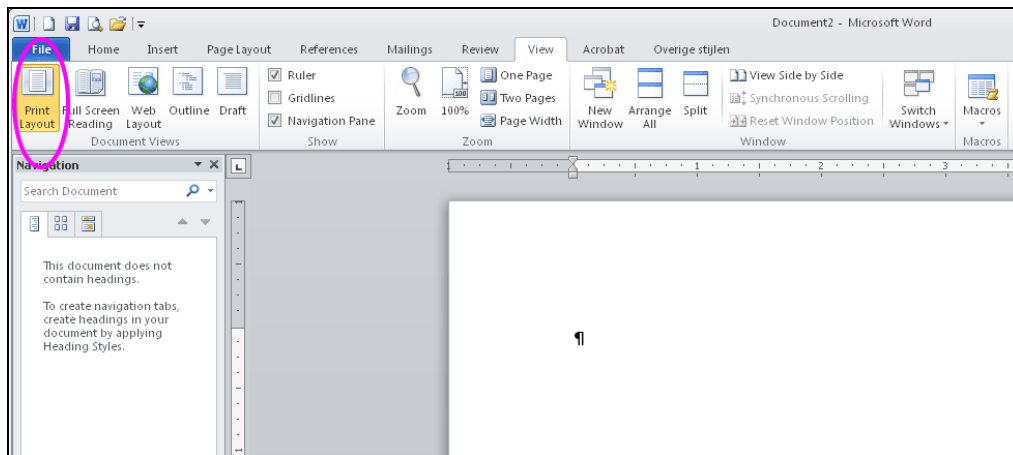


Figure 9: Print Layout in Microsoft Word.

- 8) Click on the "Home" tab, and click on the "Show/Hide paragraph marks" button (Figure 10). When this option is enabled, Microsoft Word will show all kind of normally hidden special characters, including page breaks, whitespace, tabs, and of course paragraph marks. Like with the previous step, it is possible that when you open a document the option needs to be enabled again.

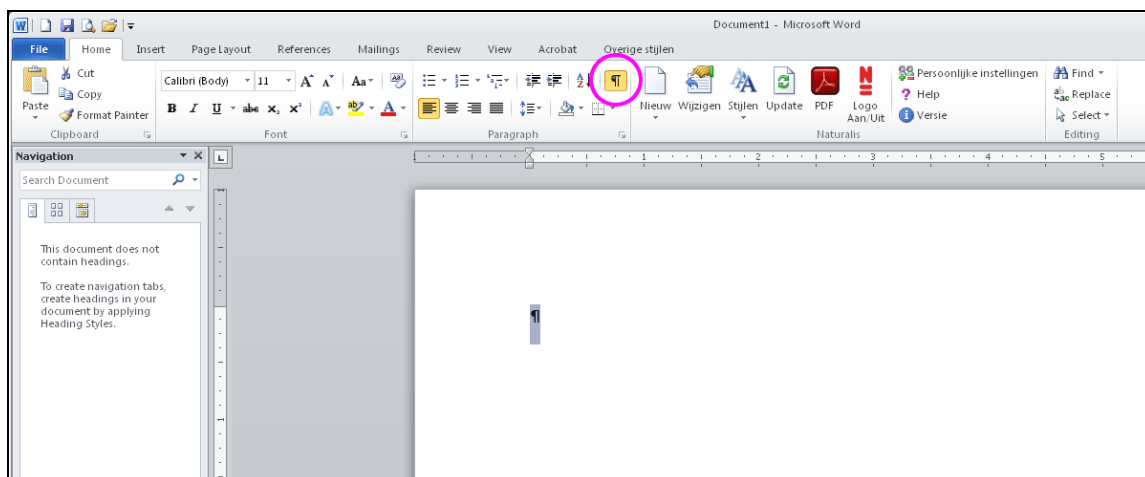


Figure 10: Enabling Microsoft Word's "Show/Hide paragraph marks" option.

The options you changed are options that you may want to change back for other work. In that case, refer back to the instructions above once you are done with preparing your legacy taxonomic text and undo the changes you made to the options.

(continued next page)

Notepad++ set up

Go to Notepad++'s "View" menu and switch on the following three options (see also Figure 11):

- "Word wrap", which wraps the text when a line is longer than your screen is wide.
- Under "Show Symbol", choose:
 - "Show White Space and TAB"
 - "Show Indent Guide"

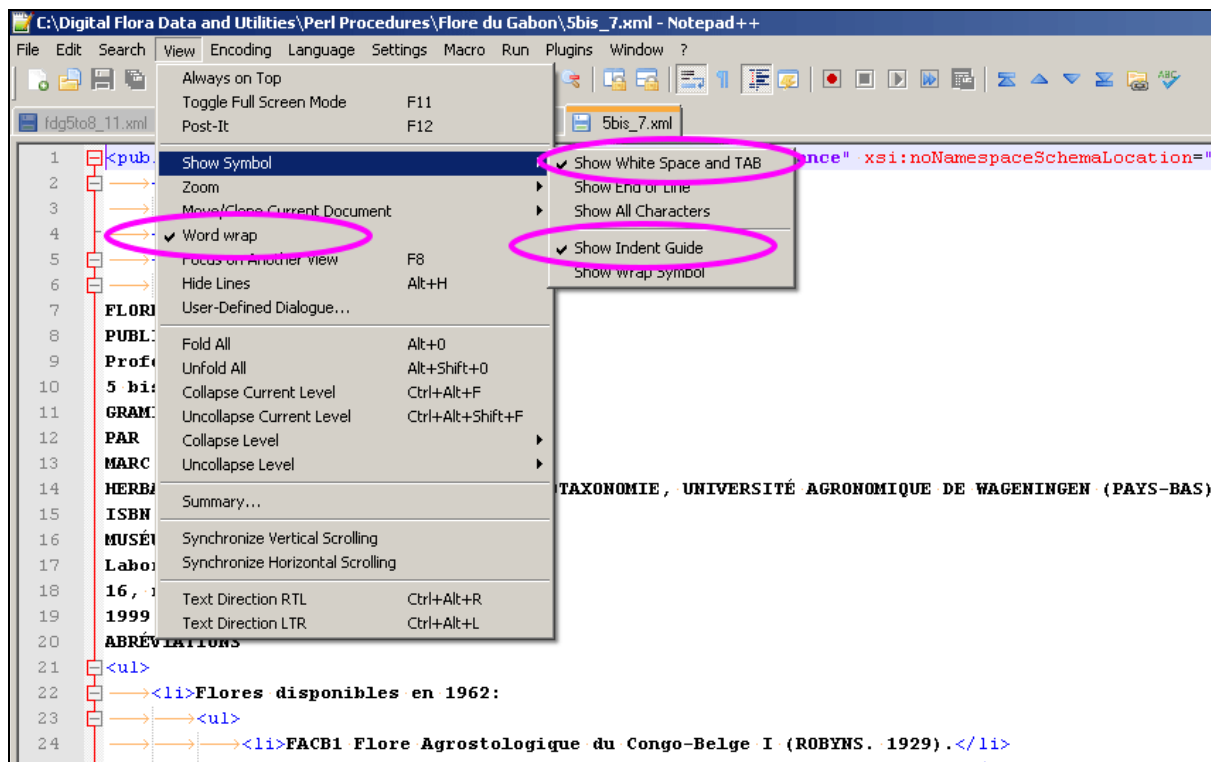


Figure 11: Helpful options in Notepad++.

Once all the programs you are going to use are properly set up, you can start working.

Extracting text from Adobe PageMaker or InDesign files

This section is optional and you should only use it if the legacy taxonomic works you are going to work with are available as Adobe PageMaker or InDesign files. Here it will be explained how to extract the text from such files, after which you can use the method described in the section titled "Cleaning up a legacy taxonomic text" to prepare the text for semi-automatic mark-up.

- 1) Start Adobe InDesign. Go to the "File"-menu and choose "Open...". Select the file(s) of your legacy taxonomic work and click the "Open"-button (Figure 12). Depending on the version of Adobe PageMaker or InDesign that was used to create the file(s) an automated conversion process may take place. If you get any warning window, read the message (it will likely be about missing fonts or broken image links) and click "OK".

Tip: If the files were made in PageMaker or InDesign on an Apple Mac, select "All files" instead of "All readable files" in the "Files of type" dropdown box in the "Open a file"-window. Then you can see the files and select them to open them.

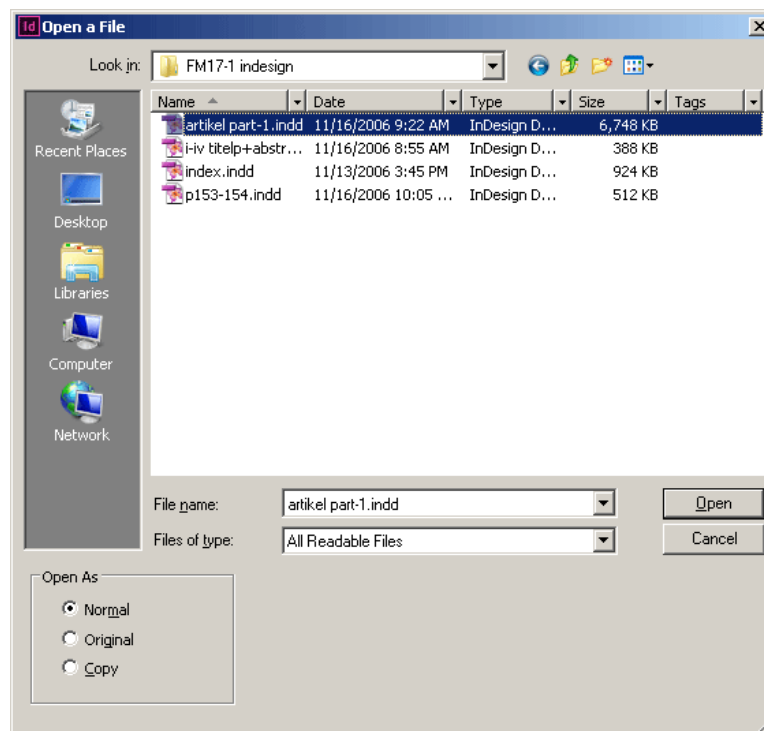


Figure 12: Opening a file in Adobe InDesign.

- 2) In Microsoft Word, start a new blank document and save it under a name you will be able to remember. Now switch back to the document you have just opened in Adobe InDesign.

In Adobe PageMaker and InDesign, text can be linked between pages and other elements on the page (e.g. text boxes), such that when something is changed on

one page or in one text box the text flow is automatically adjusted elsewhere instead of running off the page.

Depending on how the PageMaker or InDesign files were made in the past, text contained in them can be entirely linked, partially linked, or not linked at all between both individual pages and various parts of a page. Which of these is applicable to your document you will discover the moment you start selecting text.

- 3) In the first case, you can simply select all of the text and copy it to the Microsoft Word document, which is then saved. The figures do not need to be copied, but tables do. Then you can continue with the clean-up process that is described in the next section.
- 4) However, the second and third case are more work-intensive and require care to avoid skipping portions of text that have to be marked up. Figure 13 shows an InDesign document where the main text is linked together and selected, but figure captions and page headings are not. In such cases, you can best first select all the linked body text and copy it to the Microsoft Word document.

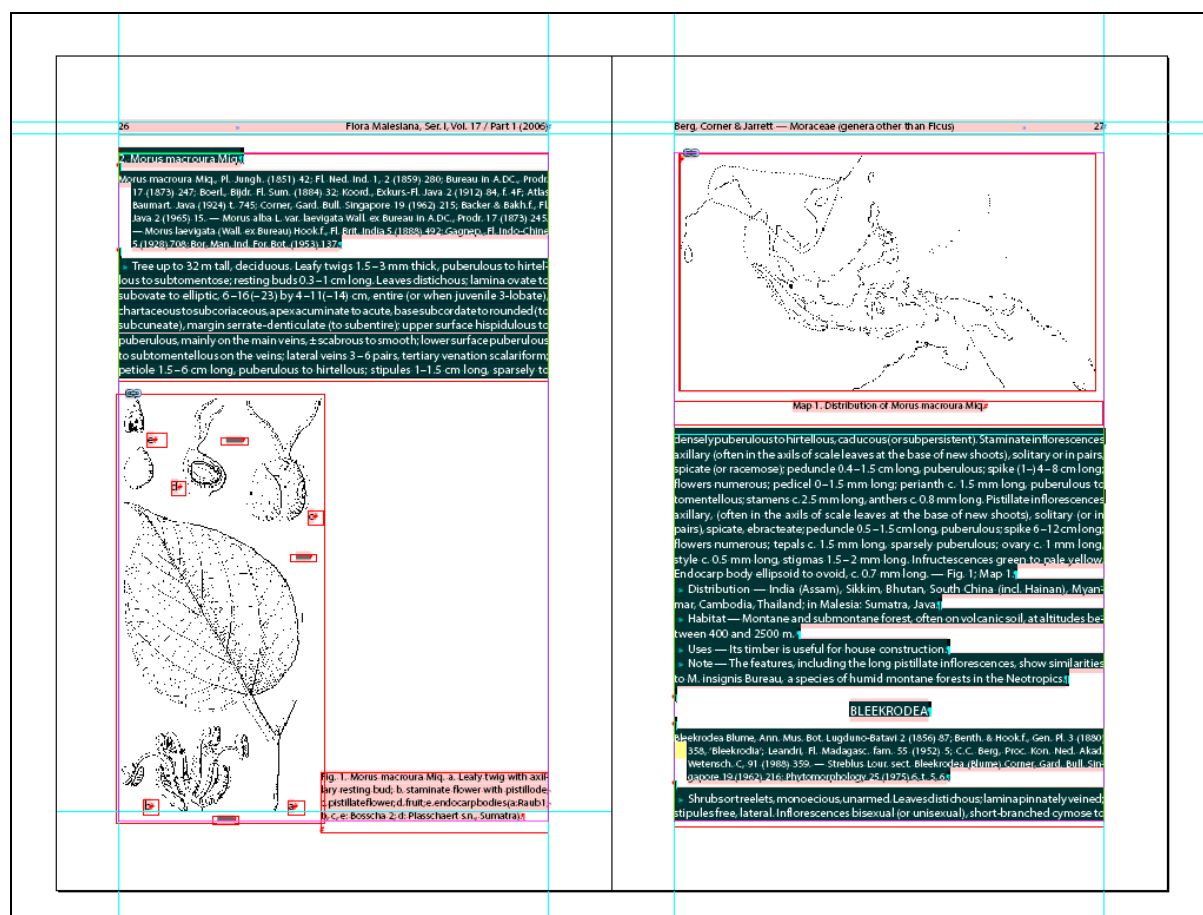


Figure 13: An InDesign document with only partially linked text.

- 5) Then you will have to copy all the text from the unlinked parts of the text to the Microsoft Word document. Try to paste footnotes near the text paragraph they belong to. Figure captions can all be pasted at the end of the taxon treatment they belong to. Figure 14 shows part of the text from Figure 13 copied into

Microsoft Word, with figure captions following the taxon treatment. The figures themselves do not have to be copied. However, you should copy tables eventually present in the text.

- a. If no text is linked at all, you will have to copy each separate text portion on each page individually. This may take some time with large documents.

2. *Morus macroura* Miq. ¶

¶

Morus macroura Miq., Pl. Jungh. (1851): 42; Fl. Ned. Ind. 1, 2 (1859): 280; Bureau in: A.D.C., Prodr. 17 (1873): 247; Boerl., Bijdr. Fl. Sum. (1884): 32; Koord., Exkurs. Fl. Java 2 (1912): 84, f. 4F; Atlas Baumart. Java (1924): t. 745; Corner, Gard. Bull. Singapore 19 (1962): 215; Backer & Bakh.f., Fl. Java 2 (1965): 15. — *Morus alba* L. var. *laevigata* Wall. ex: Bureau in: A.D.C., Prodr. 17 (1873): 245. — *Morus laevigata* (Vall.) ex: Bureau: Hook.f., Fl. Brit. India 5 (1888): 492; Gagnep., Fl. Indo-Chine 5 (1928): 708; Bor., Man. Ind. For. Bot. (1953): 137. ¶

¶

→ Tree up to 32 m tall, deciduous. *Leafy twigs* 1.5–3 mm thick, puberulous to hirtellous to subtomentose; resting buds 0.3–1 cm long. *Leaves* distichous; lamina ovate to subovate to elliptic, 6–16 (–23) by 4–11 (–14) cm, entire (or when juvenile 3-lobate), chartaceous to subcoriaceous, apex acuminate to acute, base subcordate to rounded (to subcuneate), margin serrate-denticulate (to subentire); upper surface hispidulous to puberulous, mainly on the main veins, ± scabrous to smooth; lower surface puberulous to subtomentellous on the veins; lateral veins 3–6 pairs, tertiary venation scalariform; petiole 1.5–6 cm long, puberulous to hirtellous; stipules 1–1.5 cm long, sparsely to densely puberulous to hirtellous, caducous (or subsistent). *Staminate inflorescences* axillary (often in the axils of scale leaves at the base of new shoots), solitary or in pairs, spicate (or racemose); peduncle 0.4–1.5 cm long, puberulous; spike (1–)4–8 cm long; flowers numerous; pedicel 0–1.5 mm long; perianth c. 1.5 mm long, puberulous to tomentellous; stamens c. 2.5 mm long, anthers c. 0.8 mm long. *Pistillate inflorescences* axillary (often in the axils of scale leaves at the base of new shoots), solitary (or in pairs), spicate, ebracteate; peduncle 0.5–1.5 cm long, puberulous; spike 6–12 cm long; flowers numerous; tepals c. 1.5 mm long, sparsely puberulous; ovary c. 1 mm long, style c. 0.5 mm long, stigmas 1.5–2 mm long. *Infructescences* green to pale yellow. *Endocarp body* ellipsoid to ovoid, c. 0.7 mm long. — **Fig. 1; Map 1.** ¶

→ Distribution — India (Assam), Sikkim, Bhutan, South China (incl. Hainan), Myanmar, Cambodia, Thailand; in *Malesia*: Sumatra, Java. ¶

→ Habitat — Montane and submontane forest, often on volcanic soil, at altitudes between 400 and 2500 m. ¶

→ Uses — Its timber is useful for house construction. ¶

→ Note — The features, including the long pistillate inflorescences, show similarities to *M. insignis* Bureau, a species of humid montane forests in the Neotropics. ¶

Fig. 1. *Morus macroura* Miq. a. Leafy twig with axillary resting bud; b. staminate flower with pistillode; c. pistillate flower; d. fruit; e. endocarp bodies (a: *Raubt.*; b, c, e: *Bosscha 2*; d: *Plasschaert s.n.*, Sumatra). ¶

Map 1. Distribution of *Morus macroura* Miq. ¶

¶

¶

Figure 14: Text copied from the InDesign document of figure 13.

- 6) Page headers (Figure 15) and footers do not have to be copied, as they are not used during the mark-up process. Of course, if they contain text that is of importance to the rest of the document, such as footnotes (see Footnotes using Word's footnote feature), they should be copied. Likewise, page indexes at the end of a taxonomic work can be skipped.

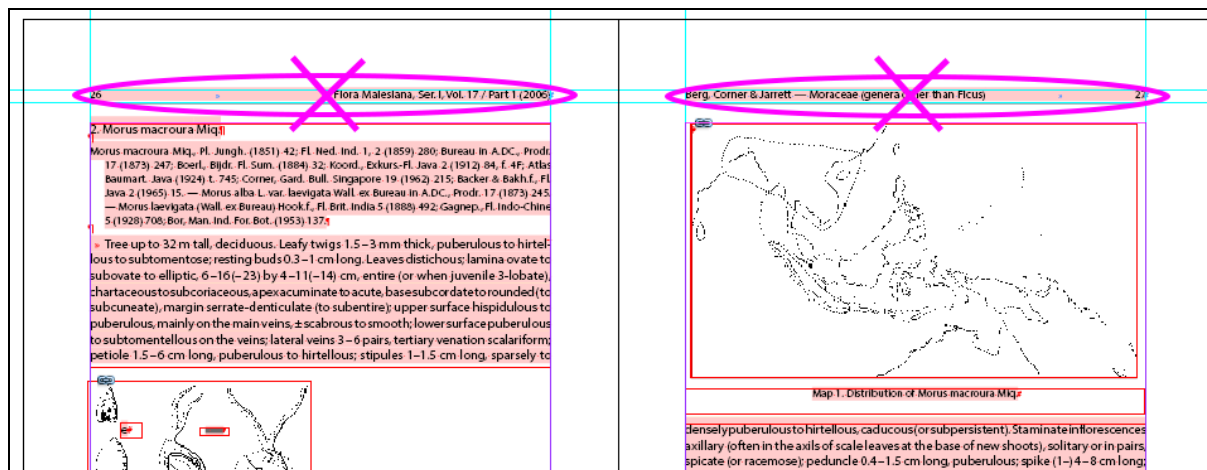


Figure 15: Page headers do not need to be copied.

- 7) Because the work is more time-intensive when not all text is linked, it is recommended that you save often. Do not forget to save when you are done. Then go to the next section.

The text extracted from Adobe PageMaker and InDesign files is going to be cleaner than OCR'ed text, so not all the points discussed in the section titled “Cleaning up a legacy taxonomic text” will apply in this case.

Extracting text from PDFs

This section is optional and you should only use it if the taxonomic works you are going to work with are available as PDFs. Here two methods will be shortly explained:

- How to extract the text from such files if they have a good quality OCR.
- How to export PDFs to Word format if they lack a good quality OCR or have no OCR at all.

After this you can use the method described in the section titled “Cleaning up a legacy taxonomic text” to prepare the text for semi-automatic mark-up.

PDFs with good-quality OCR

This section explains how to extract the text out of a PDF with good-quality OCR.

There are two options to get the text out of the PDF.

Firstly, you can save the PDF document in Microsoft Word format. Whether this successfully results in a suitable Word document depends on the quality of the OCR and, unfortunately, on the version of Adobe Acrobat Professional. Indeed, some versions of Adobe Acrobat Professional insist on running Acrobat's own very mediocre OCR on a document when saving it in Word format, ignoring any OCR that was already present. Disabling the setting responsible for this results in a Word file

without text. Adobe Acrobat Professional 7.0 does this correctly, but Adobe Acrobat Pro X (a later version!) does not.

The steps below explain how to save a PDF in Word format in Adobe Acrobat Professional:

- 1) Start by opening the PDF from which you want to extract text in Adobe Acrobat Professional.
- 2) Go to the "File"-menu and select "Save As...". Another menu opens. There, select "Microsoft Word". Yet another menu opens; in this menu select "Word Document". Figure 16 shows all menus.

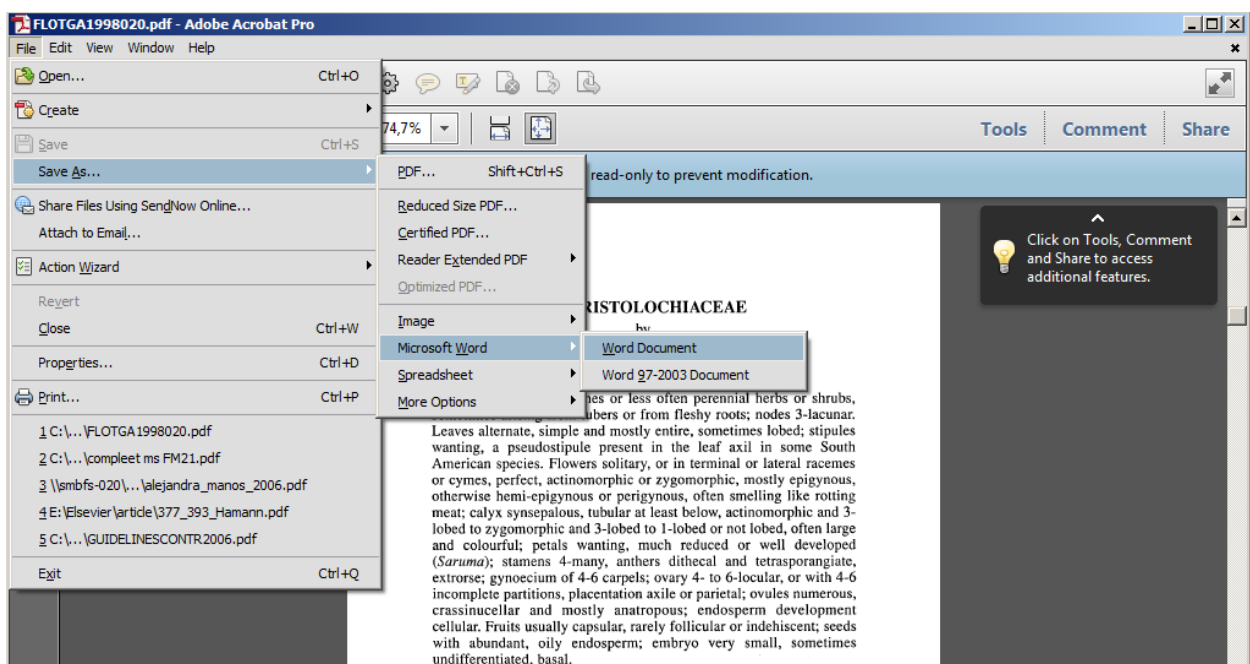


Figure 16: Preparing to save a PDF as a Microsoft Word file.

- 3) The "Save As"-window opens. In this window, select where you want to save the Word file, and click on the "Save"-button (Figure 17). The PDF will now be saved as a Microsoft Word file. This takes a few minutes, depending on the size of the PDF.

Note: If you are using a version of Adobe Acrobat Professional that forces its own OCR on the document while ignoring any OCR already present, Adobe Acrobat will also indicate it is running its OCR while saving the file. In this case, try the second option explained below.

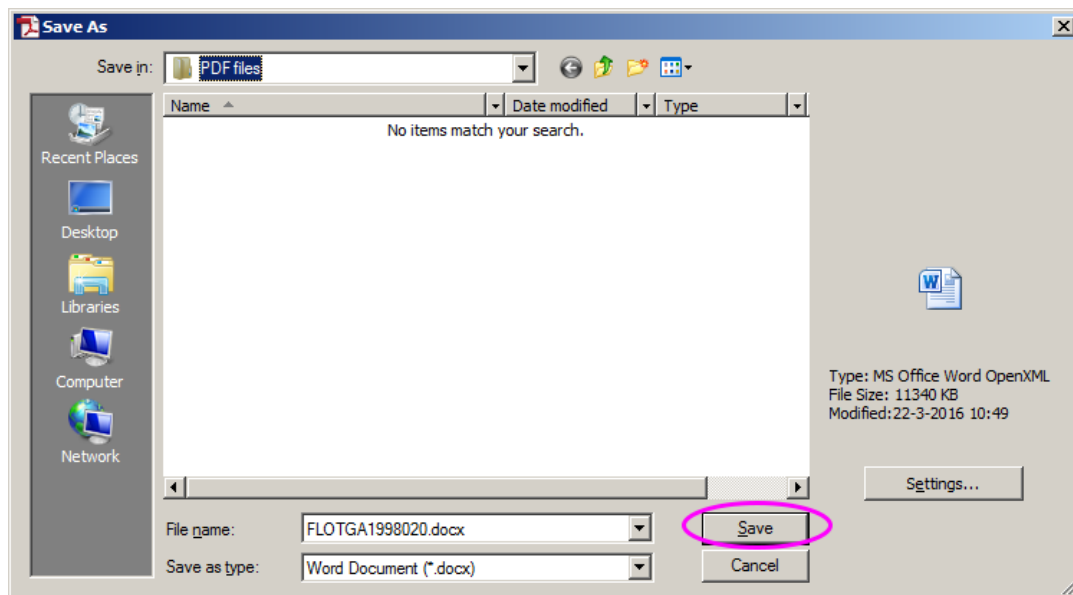


Figure 17: The "Save As"-window.

If you obtained a good quality Microsoft Word file at this point, you can continue with the section titled "Cleaning up a legacy taxonomic text". Should it still become apparent that there are problems with the Word file, you can first try the second option described below, or alternatively try to get a better OCR, as described in the next section.

The second option to get text from a PDF with good-quality OCR is to select the text in the PDF and copy it to an empty Word document.

This procedure is similar to copying text from an Adobe InDesign document, including the advantages and disadvantages. The main differences are that you work from Adobe Acrobat Professional, that there are no text frames, and that in this program you use the "Selection Tool for Text and Images" for text selection (Figure 18). There might be some other minor differences with regards to how Acrobat works, but they should not impede on your ability to copy the text to a Microsoft Word document. However, during text selection you should be attentive to any text that either cannot be selected, or is selected in an unusual order. The latter may result in messed up text once it has been copied over to Microsoft Word.

At this point you can continue with the section titled "Cleaning up a legacy taxonomic text". Should it become apparent that there are major problems with the text in the Word file, you can try to get a better OCR, as described in the next section.

(continued next page)

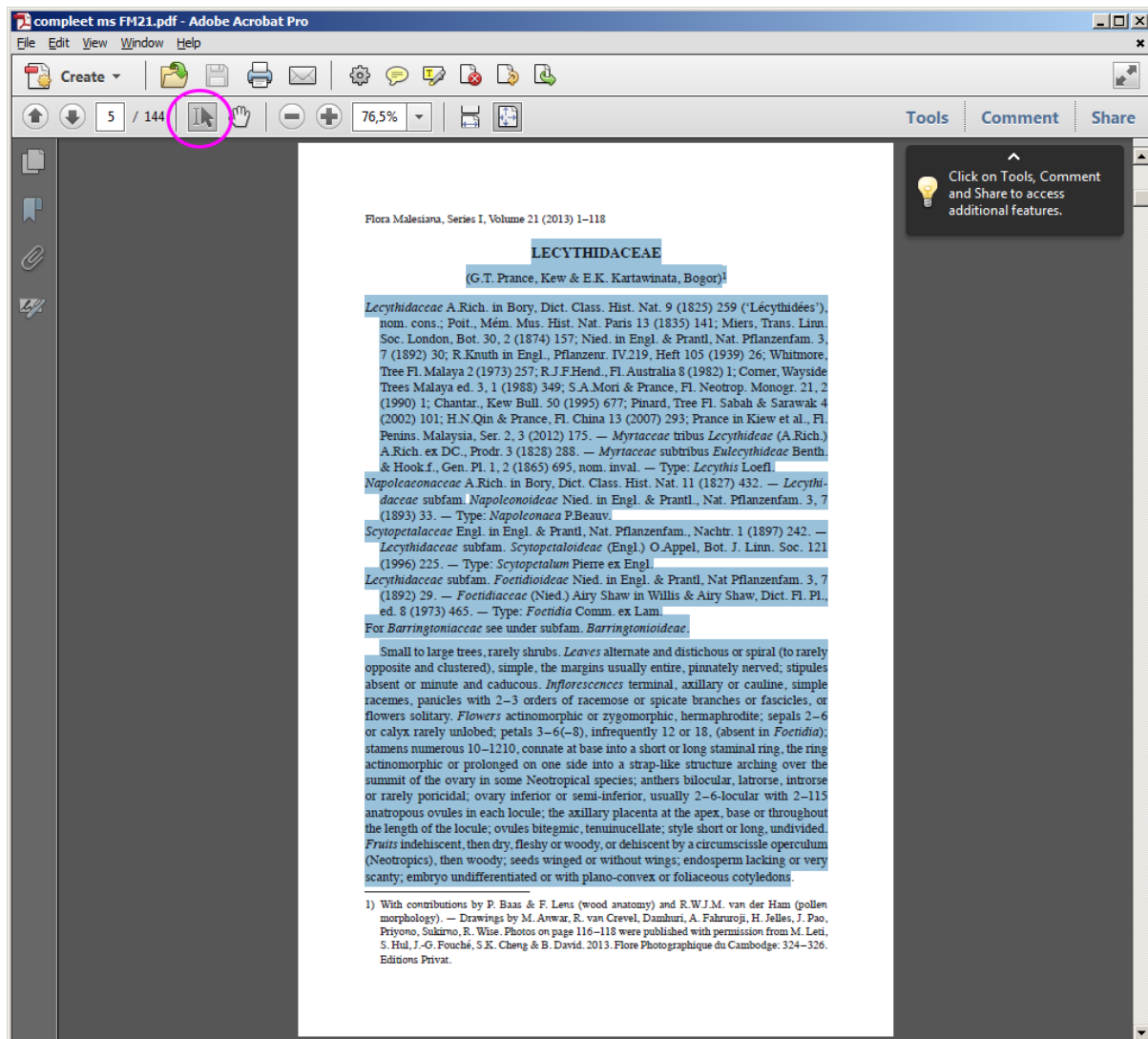


Figure 18: Selecting text in Adobe Acrobat Professional.

PDFs lacking a good quality OCR or having no OCR at all

If you only have access to PDFs that either lack a good quality OCR (many wrongly recognised characters and/or special symbols not recognised) or have no OCR at all, you will have to use special software to recognise the text in the PDF and, if possible, also export the text to Microsoft Word format.

It is suggested that you do **not** use the OCR feature of Adobe Acrobat Professional, as it is rather basic and does not support documents that contain text featuring multiple languages at once and/or with special, non-standard symbols. Indeed, using the OCR-feature of Adobe Acrobat Pro may very well be the cause of many a bad OCR in the first place.

Instead, you should use professional OCR software to recognise text in PDFs or image files, for example ABBYY FineReader Professional 12 (Academic license under €100). Short instructions for recognising text in a PDF or a series of images and exporting the result to a Microsoft Word file using this program are given below:

- 1) Start ABBYY FineReader Professional. In the "Task"-window that appears, click on "Image or PDF File to Microsoft Word" (Figure 19). The "Open Image"-window appears.

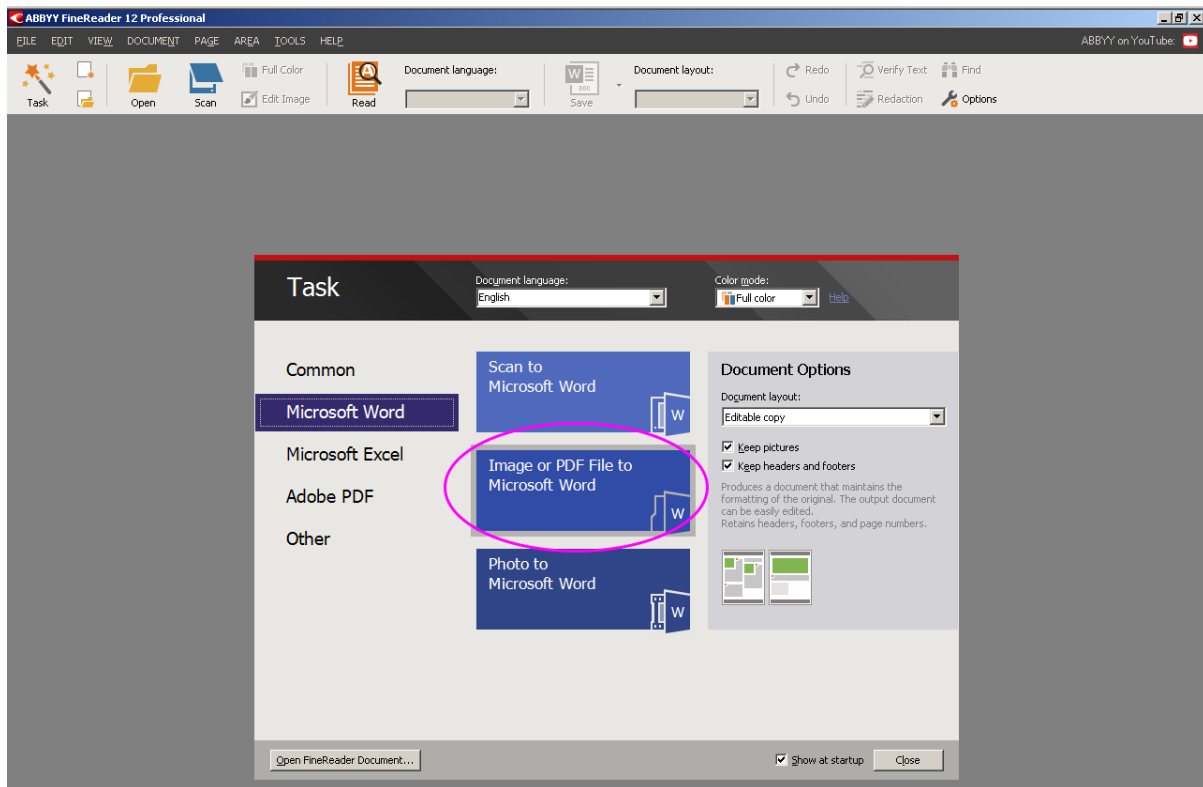


Figure 19: Selecting the correct task in ABBY FineReader Professional.

- 2) In the "Open Image"-window, choose the PDF file or series of images that you want to use. Then click on the "Options..."-button (Figure 20, next page).
- 3) The "Options"-window opens (Figure 21). In this window, go to the "Document"-tab.

(continued next page)

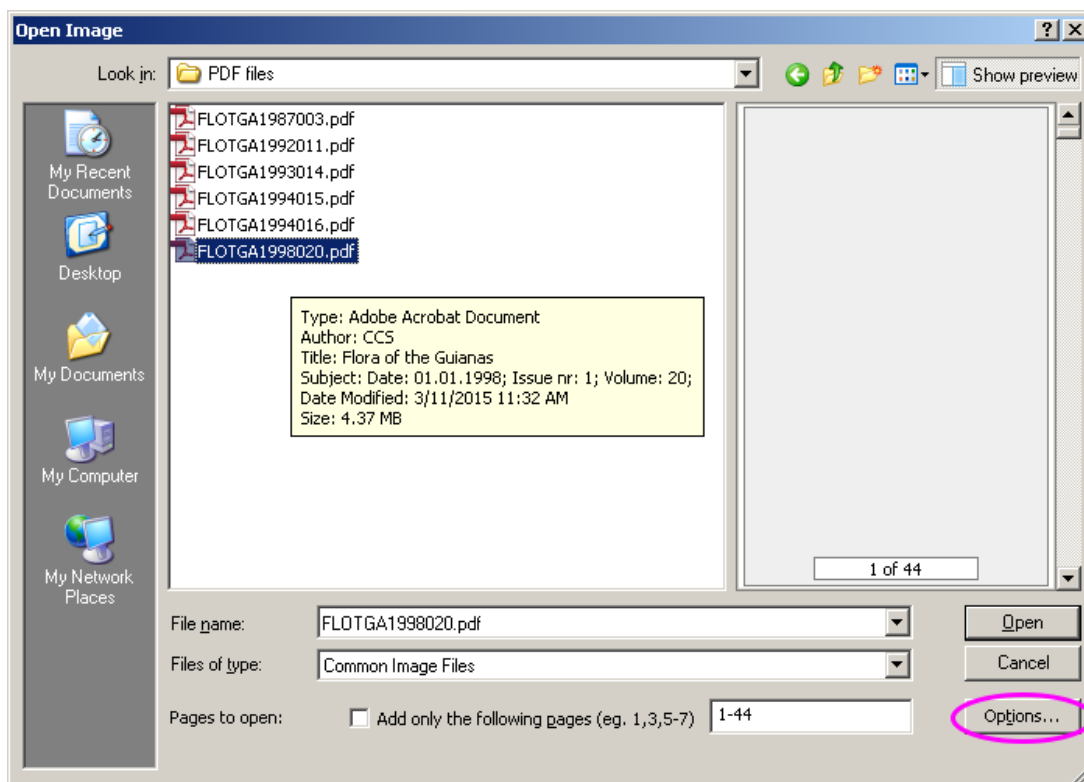


Figure 20: "Open image"-window in ABBYY FineReader Professional, with a PDF file selected.

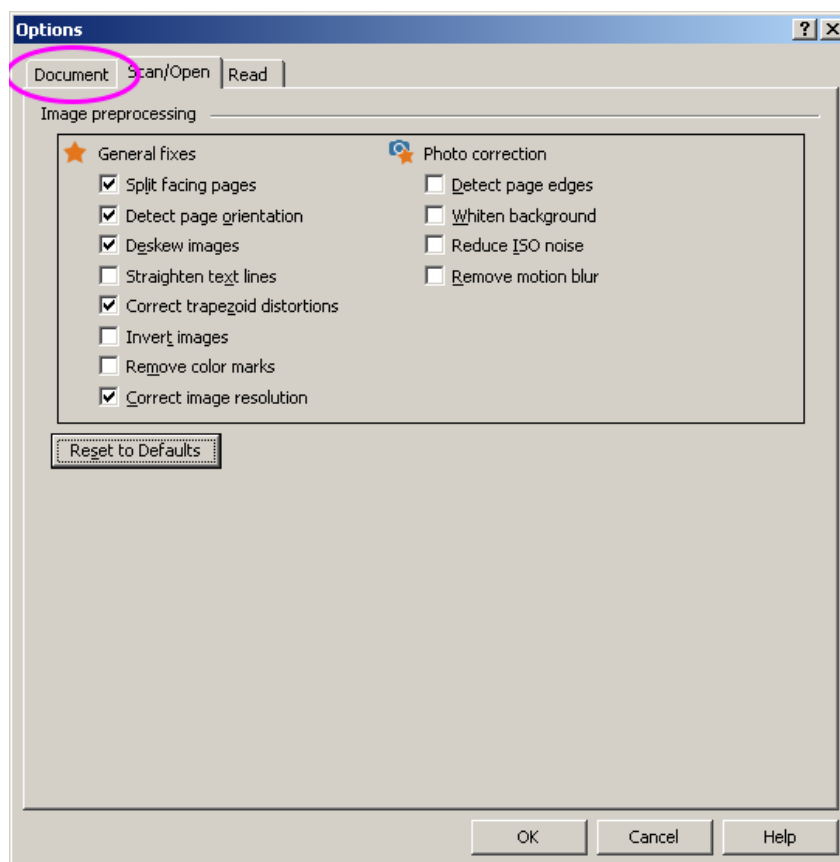


Figure 21: "Options"-window in ABBYY FineReader Professional.

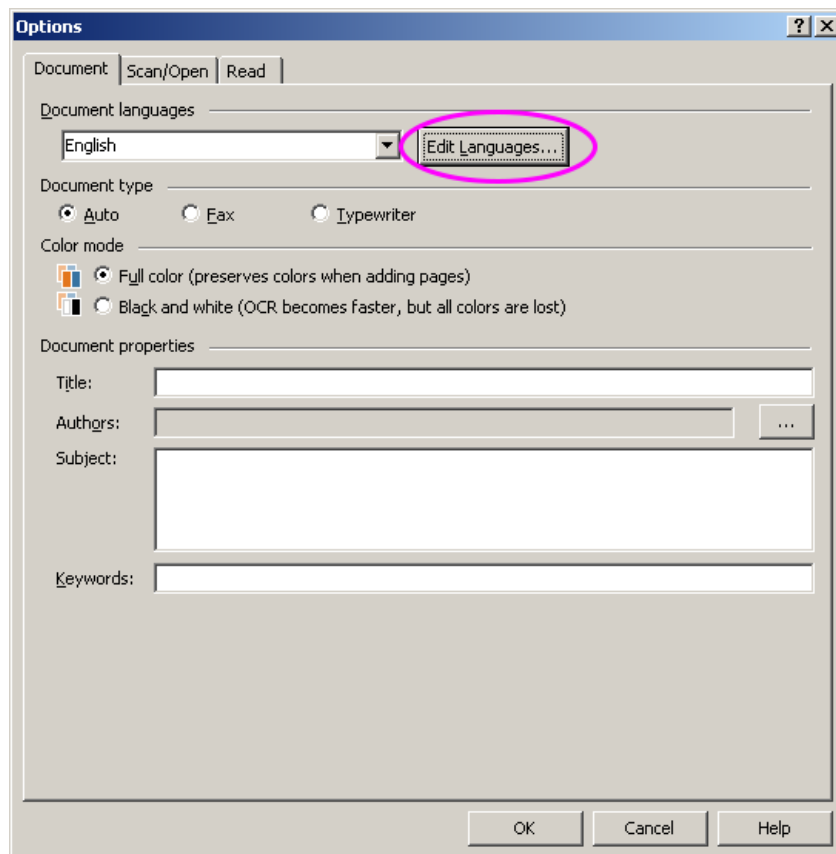


Figure 22: Document tab in the "Options"-window, with "Edit Languages..."-button circled.

- 4) In the "Language Editor"-window, choose the option "Select languages manually" (this is the default option, normally). Then, select the languages that you expect to be present in the document (Figure 23, next page). Be sure to scroll down and also have a look at the languages listed under "Formal languages", as some of them may be applicable to your document; the list can be expanded by clicking on the plus sign in front of it.
- 5) When done, click on the "OK"-button.

(continued next page)

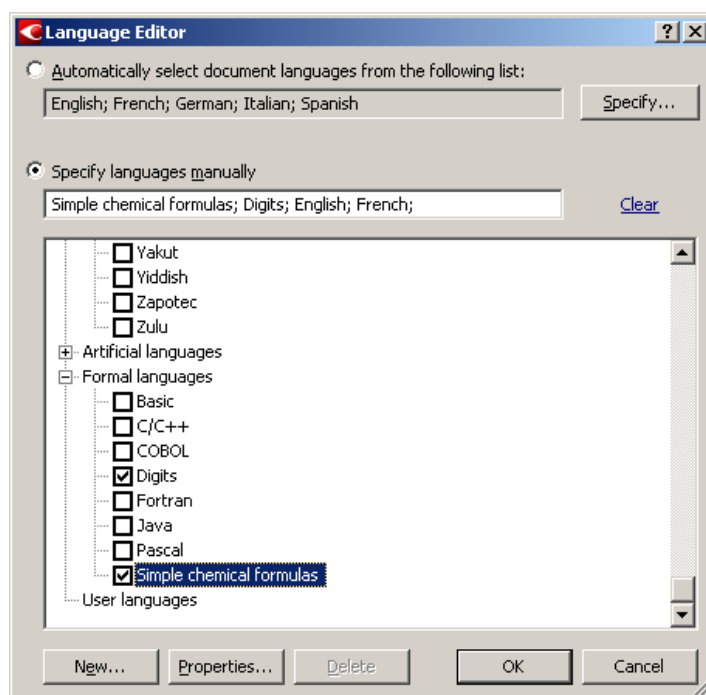


Figure 23: "Language Editor"-window in ABBY FineReader Professional.

- 6) Click on the "OK"-button in the "Options"-window.
- 7) Click on the "Open"-button in the "Open Image"-window.
- 8) Now ABBYY FineReader starts the text recognition process. This may take a while, depending on the size of the PDF. ABBYY FineReader will list eventual problems, including a list of pages with languages it does not recognise (e.g. Latin). It will also output a Word file and open it in Microsoft Word.

Generally, if you have selected the correct languages, the result will be of good to very good quality, and you can immediately continue with the next section, "Cleaning up a legacy taxonomic text".

Professional OCR software like ABBYY FineReader Professional can also learn from user corrections to further increase the accuracy of its text recognition. It may be useful if you explore the abilities of such software, especially if the source documents you are working with are from before the computer age or contain many exotic symbols such as fractions or gender symbols.

Cleaning up a legacy taxonomic text

Safeguarding the original file

It is practical to avoid working in the original Word or RTF file containing the OCR'ed text, because you may want to revert to that original file should something go catastrophically wrong during clean up.

Therefore, after you have opened the file containing the legacy taxonomic work that you are going to mark-up in Microsoft Word, go to the "File" tab, and click on "Save as" and save it under a new name that is easy to remember on a location that is also easy to remember.

For example, the file that will be used in a lot of examples in this document is saved in a folder called "fichiers nettoyes" (French for "cleaned files") under a file name that uses an abbreviation for the flora name (Flore du Gabon) and volume (Figure 24).

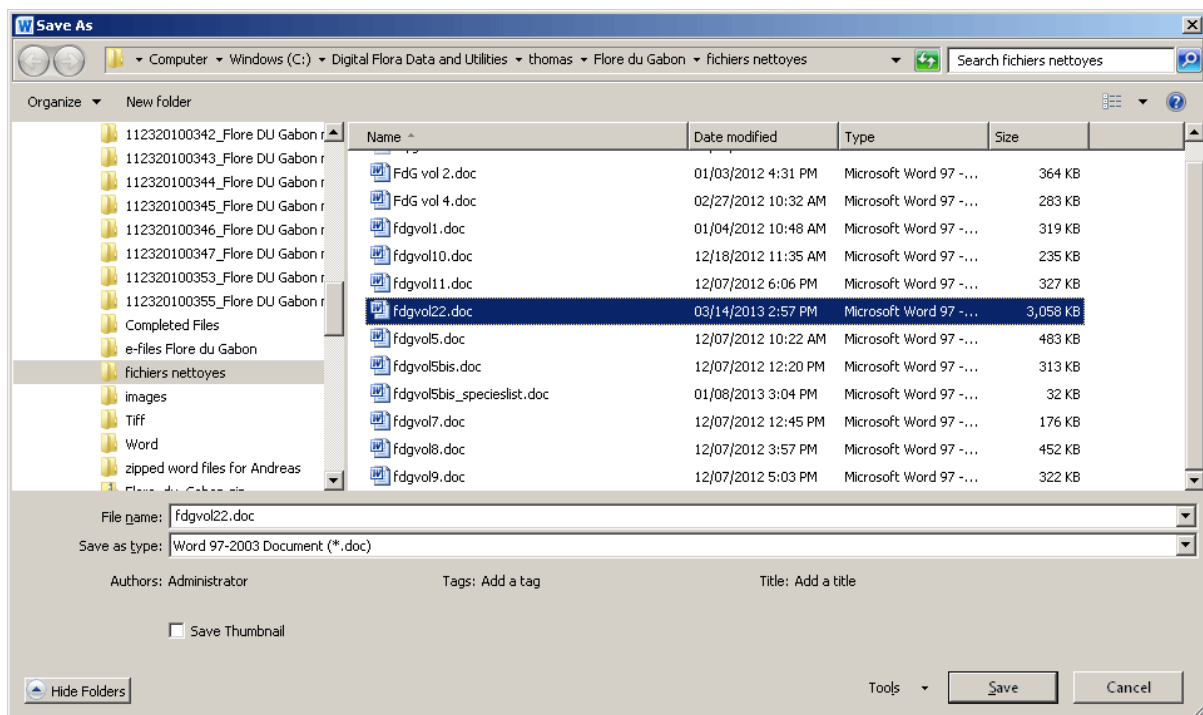


Figure 24: File containing the cleaned up Flore du Gabon volume 22.

When the clean-up process is almost finished, the file will be saved as another file with a different file type. This will be explained later on (see Saving the cleaned up text file for processing with Perl scripts).

Actual clean-up process

Overview

Cleaning up a legacy taxonomic work aims at preparing it for semi-automated mark-up with Perl scripts. This is achieved by fixing issues that are likely to interfere during the semi-automated mark-up process and are impossible to automatically fix using a script, while also removing contents that will not be required during and after the mark-up process.

In practice cleaning up a legacy taxonomic work means you work your way through the text and ensure that the text flow is not interrupted on unusual places but is interrupted in the proper places, that footnotes and figure captions are placed correctly, that all text takes up a single column, and that very specific types of contents are removed.

This section tells you what needs to be done, and what kind of issues you may encounter and how to deal with them. Everything is illustrated with figures explaining the various things you should keep an eye on and their solutions. It is suggested that you first read this section, and only then start the clean-up work, referencing back whenever needed.

Tip: Using keyboard shortcuts can seriously speed up your work. The keyboard shortcuts indicated in the instructions almost always involve first pressing down the "Ctrl"-key, keeping it down, and then pressing another key, before releasing both keys. Below is a short list of the most used shortcuts followed by what they do:

- "Ctrl-S" To save the document.
- "Ctrl-Z" To undo the previous action.
- "Ctrl-C" To copy something (text, etc.).
- "Ctrl-X" To cut something (text, etc.) from the document.
- "Ctrl-V" To paste the previously cut or copied item in a certain place
- "Ctrl-A" To select all text.
- "Ctrl-Q" To quickly remove most styles from a piece of text.
- "Ctrl-H" To quickly access the advanced search window.

Finding the various problems described below is really a learning process; you get better at it the longer you work at fixing them. Do not be afraid to try out things! You can always undo your changes (as long as you do not exit Microsoft Word).

You should read through the text *before* getting started to get an impression of the kind of problems present. When you then start cleaning up a document you can use this manual as a reference guide whenever you think you have found a problem. It is likely that you will not always encounter all of the problems identified, so do not worry if you cannot do everything that is being described.

There is a certain processing order to always keep in mind. This is described after all the potential problems and you should read it before starting (see Processing order within the document).

What to remove or fix?

The content that has to be removed or fixed usually falls into four categories:

- addenda/errata/corrigenda
- graphics
- text that is unnecessary
- text formatting that interferes with mark-up

These will be explained below.

Addenda, corrigenda, and emendanda

Sometimes, legacy taxonomic texts may contain sections with addenda, corrigenda and/or emendanda before or after the indexes at the back of a volume. These may concern the current volume, or prior volumes, or both. Usually, the volume and page each change applies to is indicated clearly (Figure 25, next page). However, instructions on what the change should be may vary from one legacy taxonomic work to another.

It is suggested that you apply these changes to the text prior to its clean-up, especially as they sometimes involve replacing entire keys or large sections of nomenclature, or adding many new taxa. If you cannot figure out what needs to be changed or have other problems, consultation of an experienced taxonomist with an affinity for the concerned legacy taxonomic work should be considered.

(continued next page)

ADDENDA, CORRIGENDA ET EMENDANDA

C. G. G. J. VAN STEENIS, *c. s.*

As was done in the preceding volumes, it seemed useful to correct some errors which have crept into the text of volumes 4–7 as well as to add additional data, new records and references to new species which came to my knowledge and are worth recording. Also there are alternative opinions about generic and specific delimitation on most of which comments are given.

Printing errors have only been corrected if they might give rise to confusion.

Volume and page number are separated by a colon. Page numbers provided with either *a* or *b* denote the left and right columns of a page respectively.

Aceraceae

- 4: 3, In Reinwardtia 7 (1965) 142 KOSTER-
592a; MANS published a new combination *Acer*
6: 915a *caesium* (REINW. ex BL.) KOSTERMANS (as
typified by *Laurus caesia* REINW. ex BL.
Bijdr. (1825) 553) to replace *Acer lauri-*
num HASSK. (cf. Fl. Mal. I, 4, 1954, 592).
The latter (earlier known as *A. niveum*
BL.) is the proper name, as the combina-
tion *A. caesium* (BL.) KOSTERMANS is ille-
gitimate because of *A. caesium* BRANDIS,
For. Fl. (1874) 111, Atlas t. 21.
Unfortunately this was overlooked by
WHITMORE, Tree Fl. Malaya 2 (1973) 1.

Amaranthaceae

- 4: 73; *Celosia argentea* L. var. *cristata*.
5: 554a A biosystematical study by Dr T. N.
KHOSHOO (Bull. Bot. Surv. India 12, 1970,
67–69, 1 fig., 2 pl., 1972) has shown that
C. argentea must be the ancestral form
from which var. *cristata* must be derived.
4: 86b C. C. TOWNSEND (Kew Bull. 29, 1974,
464) has transferred *Aerva curtisii* OLIV.
to a new genus *Psilotrichopsis* to accom-
modate this species and *A. cochinchinensis*
GAGN. The new genus is said to differ
from *Psilotrichum* by verrucose seed and
structure of the pollen wall, and from
Aerva besides by opposite leaves and mul-
tinerved petals.
4: 93a, For *Alternanthera bettzickiana* (REGEL)
594b; NICHOLS., which in vol. 4 was distin-
guished as a variety of *A. ficoidea* (L.) R.
BR., KANIS (Contr. Herb. Austr. 1, 1972,
6) made a new combination: *A. manillensis*
6: 916a (WALP.) KANIS. As it later appeared that
WALPERS' basionym belonged to another
species, KANIS (*ibid.* 7, 1974, 7) cancelled
this name in favour of the one accepted
in Fl. Mal. vol. 6, *l.c.*

Burmanniaceae

- 4: 17a *Burmannia coelestis* DON.
Add to synonymy: *Cryptonema malac-*
censis TURCZ. Bull. Soc. Nat. Moscou 21
(1) (1848) 590, non *Cryptonemia* AGARDH,
1842; Fl. Dahur. 1 (1848); WALP. Ann. 3
(1852) 609. — *Nephrocoelium malaccense*
TURCZ. Bull. Soc. Nat. Moscou 26 (1)

(1853) 287; Fl. Dahur. 1 (1853). —
Nephrocodum malaccense WALP. Ann. 6
(1861) 41, *sphalma*.

These three generic names should also
have been added to the synonymy of the
genus *Burmannia* L. on p. 15. Cf. JONKER,
A monograph of the Burmanniaceae.
Thesis, Utrecht (1938) 121.

Burseraceae (LEENHOUTS)

- 5: 213 *Protium* BURM. f.
Correct in Distr.: In continental Asia
there is but one species: *P. serratum*
(COLEBR.) ENGL., of which *P. yunnanense*
(HU) KALKM. is a synonym. The latter
should be (nearly) glabrous and have
somewhat larger fruits; these characters
appear to be grading, however.
5: 214b *Protium macgregorii* (F. M. BAILL.)
LEENH.
Add to references: HOOGL. in Walker
(ed.), Torres Straits Symp. (1972) 151, f.
8.21 (map).
Add to synonymy: *Dracontomelum pa-*
puanum LAUT. in K. SCH. & LAUT. Nachtr.
(1905) 301.
It occurs also in SE. New Guinea:
SCHODDE & CRAVEN 4685.
5: 222b *Dacryodes costata* (BENN.) H. J. LAM.
Add to description: Inflorescences appar-
ently sometimes exclusively axillary (SAN
75957).
5: 227a *Dacryodes macrocarpa* (KING) H. J. LAM.
Add to synonymy: *D. expansa* (non H. J.
LAM) KALKMAN, Blumea 7 (1954) 510, f.
2 a & b, *typo excl.*; LEENH. Fl. Mal. I, 5
(1956) 228, *ditto*; *ibid.* I, 6 (1972) 919.
5: 227b Replace KEY TO THE VARIETIES by the
following:
1. Leaves 4- or 5-jugate. Philippines
var. *merrillii*
1. Leaves up to 3-jugate.
2. Leaflets widest about the middle,
equal-sided at base; nerves at a right
angle to the midrib. Sarawak, Brunei
var. *patentinervia*
2. Leaflets widest in the lower half,
oblique at base; angle between mid-
rib and nerves acute.
3. Twigs and axial parts of leaves
smooth, blackish when dry; leaflets
rather thick and stiff, midrib and
nerves not sharply prominent on

(820)

Figure 25: A typical page with addenda, corrigenda, and errata for Flora Malesiana.

Graphics

Graphics in legacy taxonomic text documents in Microsoft Word include the following:

- figures and photographs
- text that was improperly recognized during the OCR may show up as a graphic

The first will have to be removed (see "Removing graphics - basics"), while in the second case the text that was not properly recognized will have to be typed out before removing the graphic. Locations where such OCR problems frequently occur are:

- a. Figures where sometimes part of the caption is recognised as part of the figure.
- b. Special symbols in text, e.g. the symbols for sexuality ($\varnothing\sigma\varnothing$ etc.).

Especially the latter are very hard to spot. Fortunately, Microsoft Word's search function can easily find graphics:

- 1) First, open the "Advanced Find..."-window by clicking the small arrow to the right of the "Find"-button on the "Home"-tab (Figure 26).

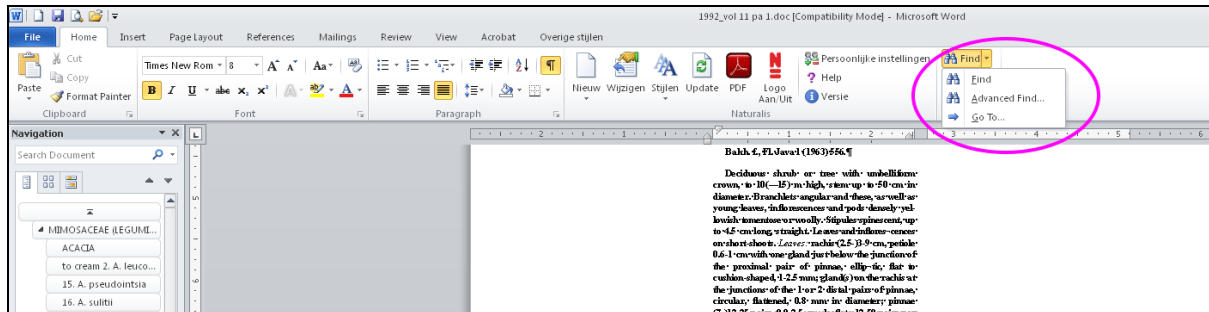


Figure 26: Starting Microsoft Word's Advanced Find function.

- 2) In the "Advanced Find"-window, click the "More >>"-button (Figure 27) to show all of the advanced options. Then click on "Special" (Figure 28).

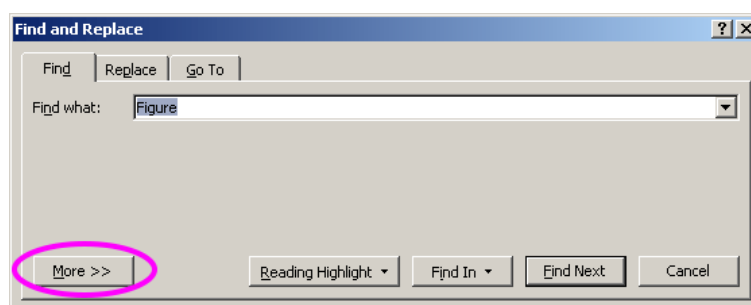


Figure 27: Advanced find window, "More >>"-button.

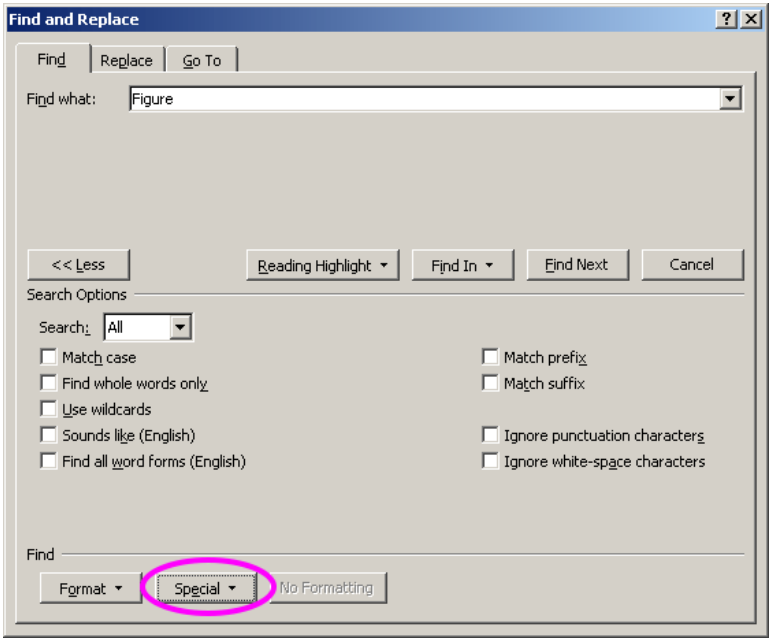


Figure 28: Advanced search options in Microsoft Word.

- 3) In the list that shows up, select "Graphic" (Figure 29). Now the "Find what:" text field will contain "^g" (Figure 30).

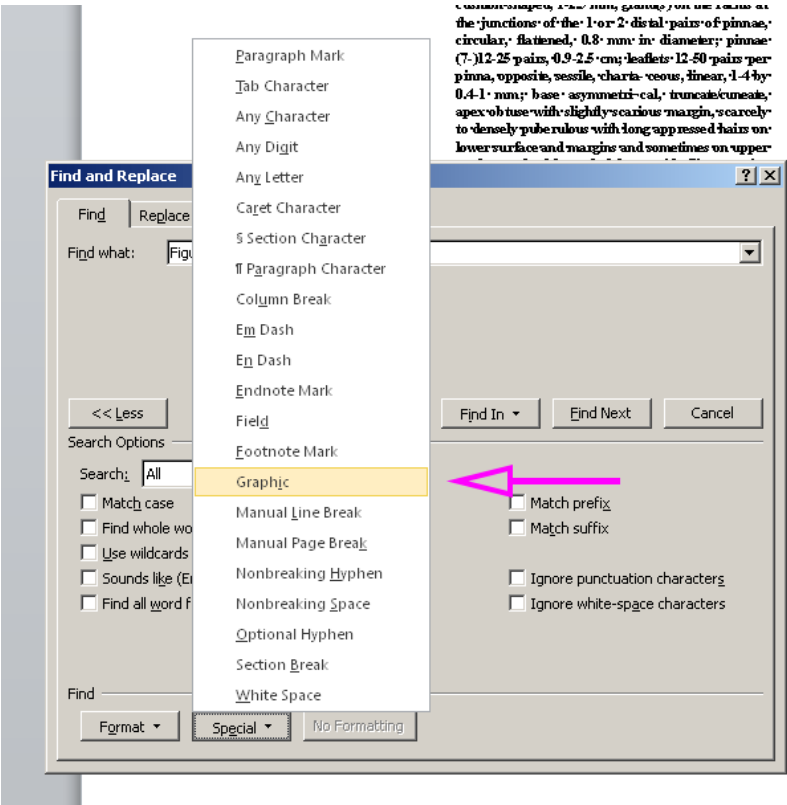


Figure 29: List of special search options in Microsoft Word.



Figure 30: Searching for graphics.

- 4) If you now click the "Find next"-button a few times, you'll notice you can now navigate easily from graphic to graphic to check for problems. You can fix text that was not properly recognized at this point (see Fixing unrecognized text), but before you try to remove graphics, keep in mind the order in which you should process the document (see Processing order within the document).

Removing graphics - basics

To remove a graphic, you do the following:

- 1) First, you select it by clicking on it once. A selection box with dotted lines appears around the graphic (Figure 31).
- 2) Then you press the "Delete"-key on your keyboard to delete the graphic.

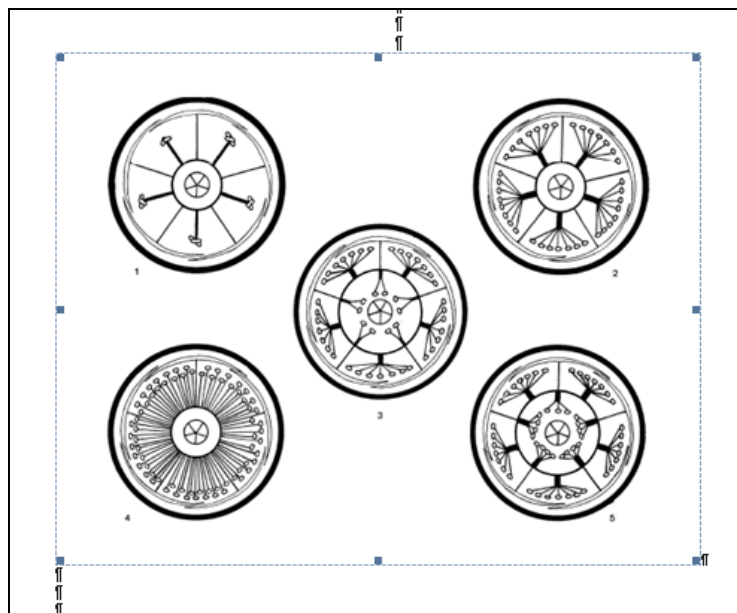


Figure 31: A selected graphic in Microsoft Word, with dotted selection box surrounding it.

Removing graphics in text boxes

In some cases the graphic is placed in a text box that may or may not contain the figure caption. If the figure caption is not part of the text box but of the main text, you can simply select the text box containing the figure and delete it.

However, if the opposite is true, the text of the caption will have to be cut from the text box and pasted in the main body text. An example of such a case is shown in Figure 32.

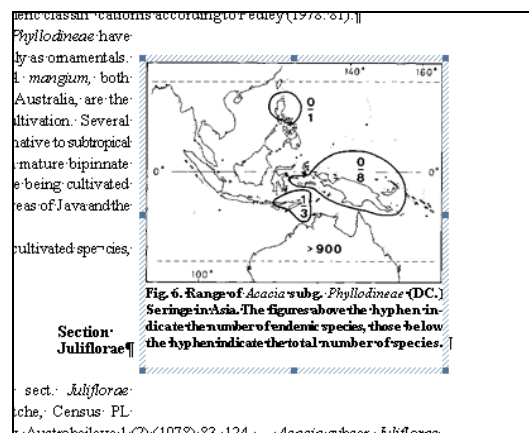


Figure 32: Figure caption in text box.

To fix this, you do the following:

- 1) In the text box, you select the figure caption using the mouse cursor (Figure 33).

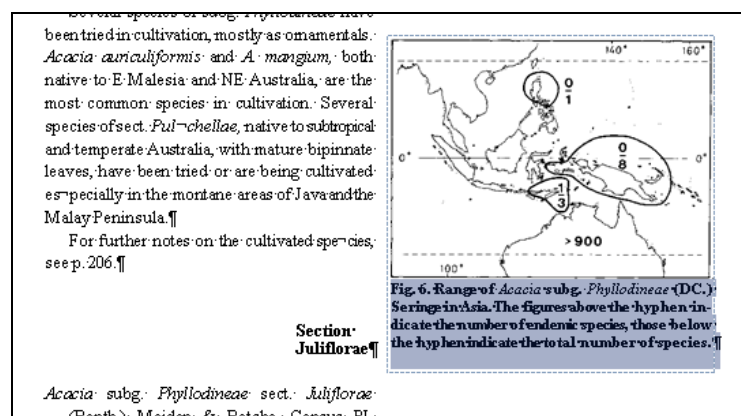


Figure 33: Selecting text in a text box.

- 2) Now you cut the text out of the text box (keyboard shortcut Ctrl-X) and paste it in the main text (keyboard shortcut Ctrl-V) (Figure 34).

(continued next page)

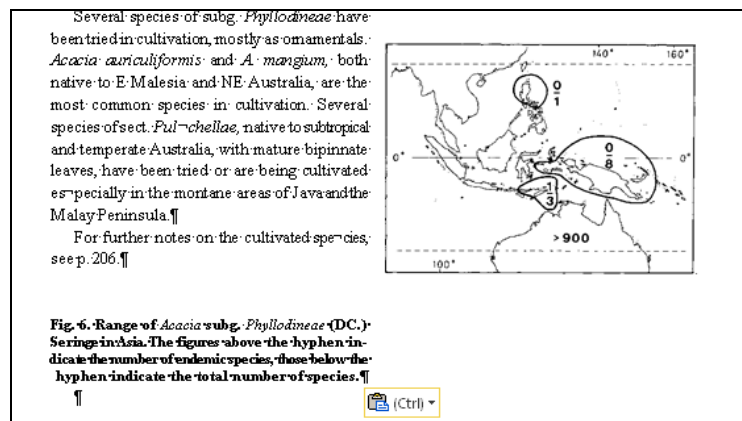


Figure 34: Figure caption is cut out of text box and pasted in main text.

3) Then you delete the graphic in the text box (Figure 35).

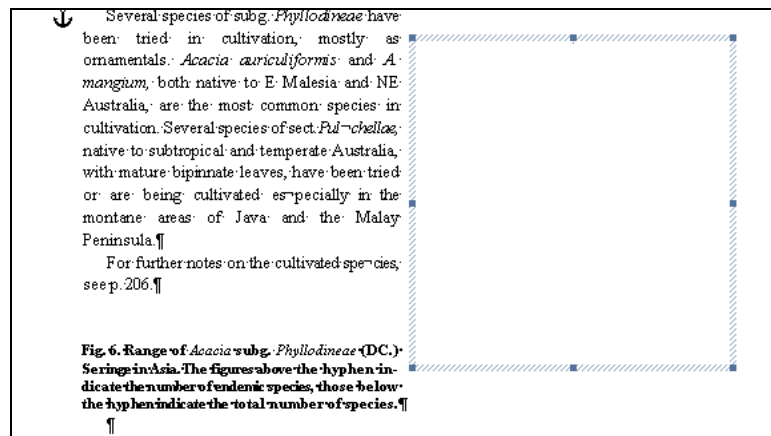


Figure 35: Deleting the figure in the text box.

4) Finally you delete the text box (Figure 36).

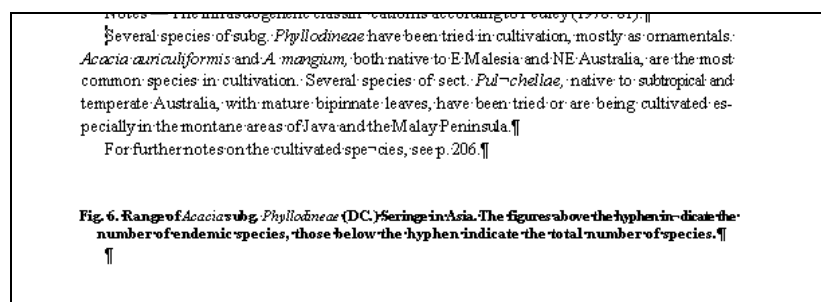


Figure 36: After deleting the text box.

Fixing unrecognized text

If text that was not properly recognized by the OCR is part of a graphic, you need to manually retype the text in the main text body (in the correct location) before deleting the graphic as explained earlier.

However, for symbols for plant sexuality, the procedure is somewhat more involved:

Figure 37 shows a wrongly recognized symbol for plant sexuality found using the search method for graphics described above.

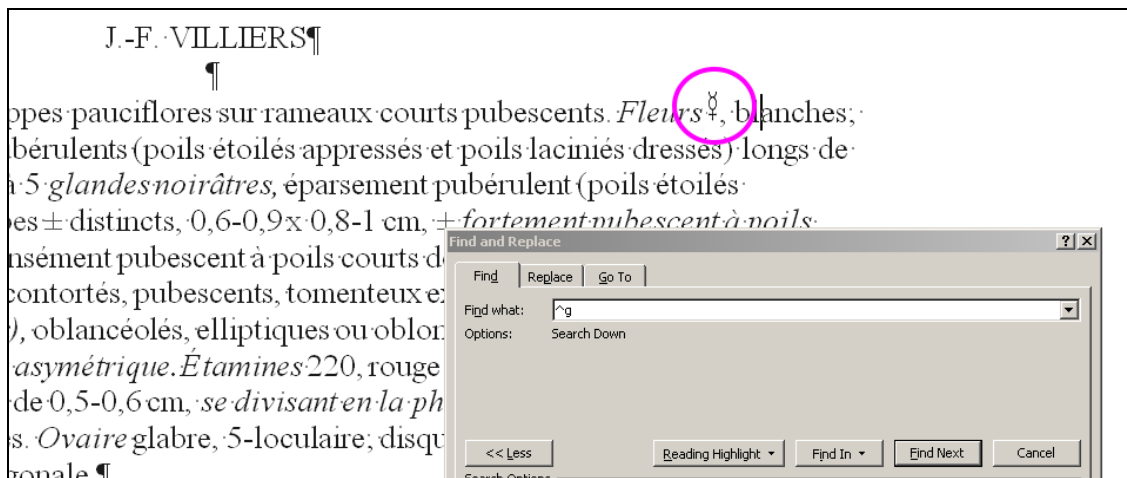


Figure 37: Wrongly recognized plant sexuality symbol in description.

One option for you to insert the proper text character instead of the graphic is to click on the "Symbol"-button on the "Insert"-tab, and then to choose "More Symbols..." (Figure 38). You then can look through dozens upon dozens of characters to try to find the character(s) you need. This is fairly tedious.

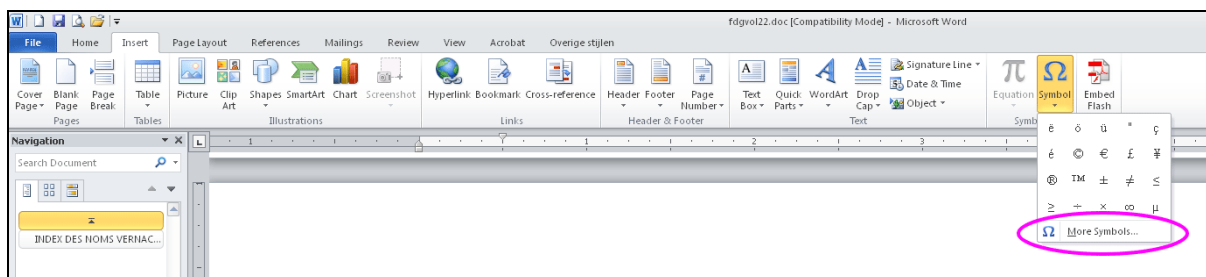


Figure 38: "More symbols..."-option to try to find and insert plant sexuality characters.

A much better option is to use the web to find the symbols you need. The following places have them:

- The Wikipedia Gender Symbol page:
http://en.wikipedia.org/wiki/Gender_symbol
- This page lists many of the special characters available in the Unicode standard: <http://www.utf8-chartable.de/unicode-utf8-table.pl> (you need to select "U25A0 ... U+25FF: Geometric Shapes" in the drop down list).

Use whichever option has the symbol you need for you. You should select the symbol you want to use, copy it, and paste it into your Microsoft Word document. A third option is to copy the symbols out of this manual.

For the example shown in Figure 37, the result after copy and pasting the correct symbol in the text is shown in Figure 39.

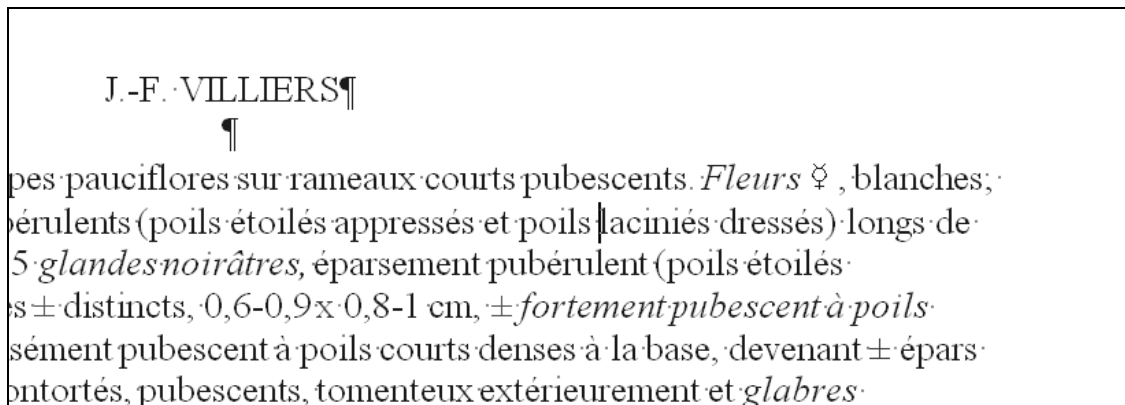


Figure 39: Proper plant sexuality symbol inserted instead of graphic.

Technical aside: The technical reason that causes some symbols not to be recognised during OCR is that symbols for plant sexuality are special characters that are not supported by all fonts. The symbols for male (♂), female (♀), and the older symbol for hermaphroditism (⚥) are all supported by many Windows fonts, but unfortunately the more modern symbol for hermaphroditism (that looks like a combination of ♂ and ♀) cannot be *displayed* in Windows - it shows up as a white rectangle: □ (this may show up correctly in the future due to better Unicode support in Windows). This is because Windows unfortunately has a fairly limited ability to display Unicode characters in most fonts compared to other operating systems. However, the character is actually present and properly identified internally in Windows; you just cannot see it as anything but a box.

Tip: If you want to be sure you copied the right symbol, paste it into your internet browser's search field (does not work in all browsers).

Preparing graphics

The Figures you have removed will have to be prepared for online use separately. This is explained in **image processing.doc**.

Unnecessary text

The transition from paper-based publications to web-based publications using XML mark-up means certain specific types of contents of legacy taxonomic works become useless.

Legacy taxonomic works often were printed using fonts that do not exist anymore today or that have divergent kerning (the distance between characters) compared to modern fonts. When legacy taxonomic works are digitized and have OCR applied to them the original font gets matched to a font that is as close to the original as possible, but there will still be differences. These differences will cause shifts in the

location of text on a page and will also cause changes to the pagination. Online versions of legacy taxonomic works will likely use fonts that facilitate the online reading experience, causing further differences from the original paper publications. As it is impossible to perfectly reproduce a legacy printed publication without recording the precise details of each page's structure, printed versions of digitized legacy taxonomic works will therefore likely have a different pagination than the original work. Furthermore, if the online version of a legacy taxonomic work is to be updated, further changes in pagination of a printed version will occur.

Therefore, conserving the pagination information of the original taxonomic work is not required. The same applies to tables of contents and indexes. A decent search engine can easily replace all of these in online versions of legacy taxonomic works, while they can be generated on the fly when producing a printed version based on the online version using a template and the mark-up inserted in the document.

In practice, the consequences of the above are that you need to remove the following from your legacy taxonomic work text:

- Page headers and footers (not footnotes).
- Tables of contents and indexes.

Headers and footers may contain the following information:

- page numbers
- author(s)
- publication title
- subject
- volume number and (eventually) part number

Headers and footers

Headers and footers can occur in several places in a document. The following instructions explain where this is and how you can remove the header and footer text in each case:

- 1) Firstly, they can be present in the actual header and footer of a Microsoft Word document (Figure 40). When this is the case, the text will be coloured grey instead of black.

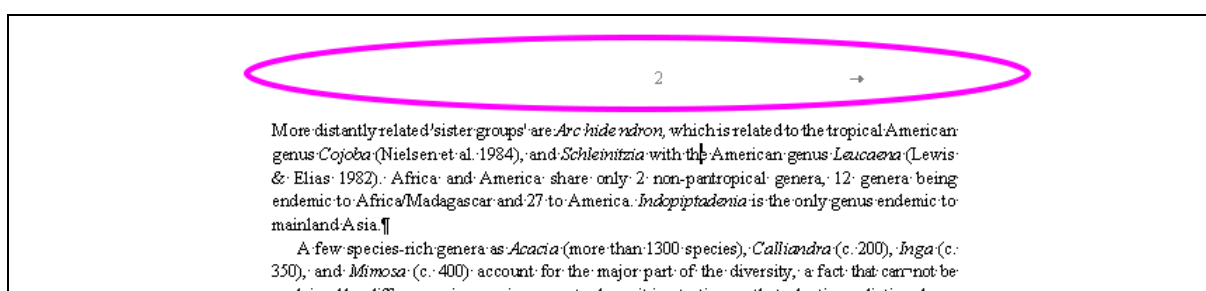


Figure 40: Page header with text contents.

- 2) In such cases, you can select the header by double-clicking on the header text. This will also make the "Header & Footer Tools"-tab appear (Figure 41).

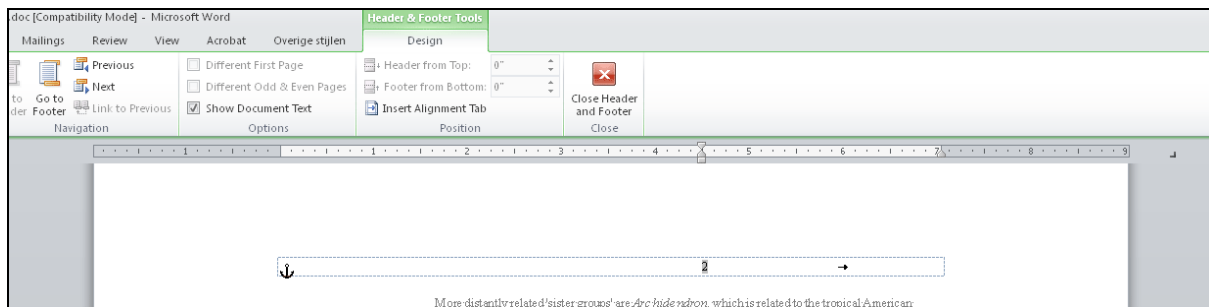


Figure 41: A selected header.

- 3) You can now select the text in the header and delete it. Then press the "Close Header and Footer"-button on the "Header & Footer Tools"-tab (Figure 42).

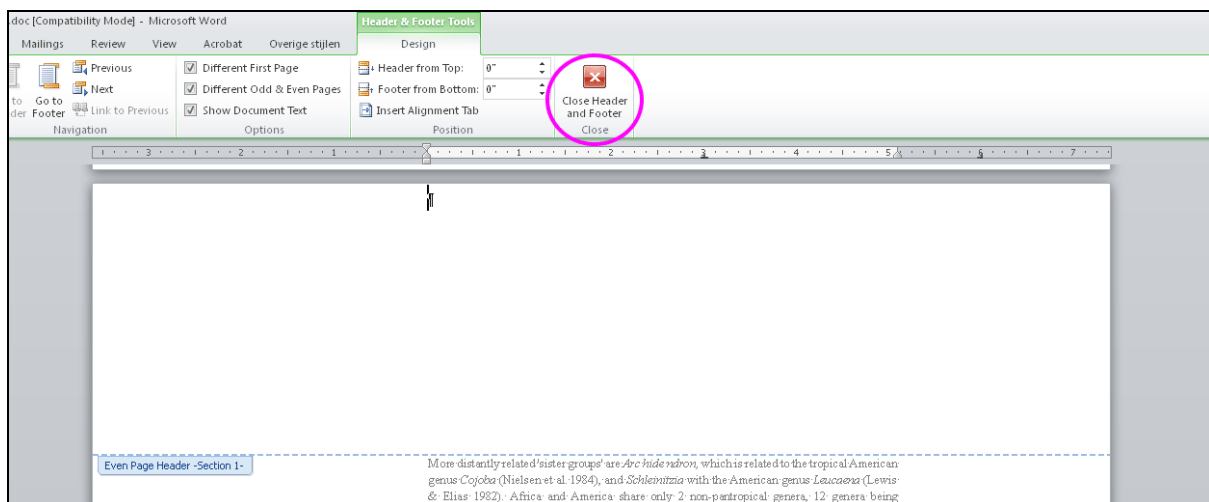


Figure 42: The header text has been removed.

- 4) Secondly, the header and footer text can be located in a text box at the top or bottom of the page (Figure 43). In this case, you just delete the text box.

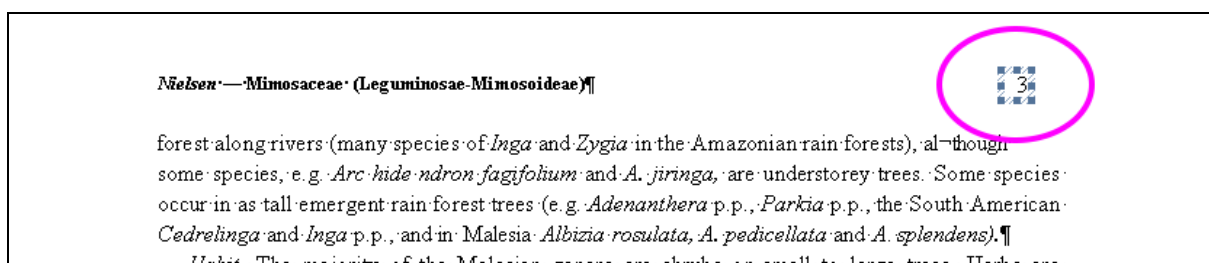


Figure 43: Header text in a text box.

- 5) Thirdly, the header or footer text can be part of the main text (Figure 44). In that case, you select the text and delete it.

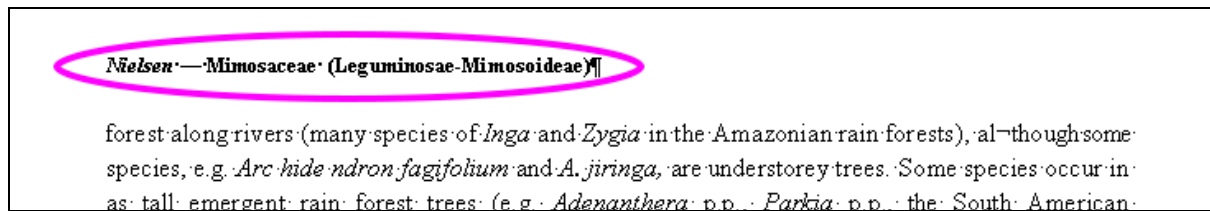


Figure 44: Header as part of the main text.

- 6) Should a footer accidentally contain a footnote, then you should move the footnote to the main text.

Tables of contents

Tables of contents are generally located at the beginning of a taxonomic text (Figure 45), but they sometimes are also located at the start of individual chapters or family treatments.

SOMMAIRE	
Celastraceae	3
Pandaceae	14
Bombacaceae	31
Cannabaceae	53
Bixaceae	59
Avicenniaceae	63
Index des noms scientifiques	67
Index des noms vernaculaires	70
Illustrations de l'auteur	
ABRÉVIATIONS	
FFSG : AUBREVILLE, Flore forestière soudano-guinéenne (1950).	
FFCI : AUBREVILLE, Flore forestière de la Côte d'Ivoire, ed. 1, 1 (1936); ed. 2, 1 (1959).	
FTA : OLIVER, Flora of tropical Africa.	
FTEA : TURILL & MILNE-REDHEAD, Flora of tropical East Africa.	
FFNR : WHITE, Forest Flora of Northern Rhodesia (1962).	
FFFT : BURTT-DAVY, A manual of flowering plants and ferns of the Transvaal and Swaziland 2 (1932).	
FWTA : HUTCHINSON & DALZIEL, Flora of West tropical Africa, ed. 1, 1 (1927-1928).	

Figure 45: A table of contents at the start of a legacy taxonomic work.

Tables of contents should only be deleted if they do not contain any taxonomic information that is not repeated elsewhere (some older taxonomic works combine their table of contents with a simplified classification tree - these are best treated as a table or a figure). To remove them, you select the text and press the "Delete"-key on your keyboard.

Indexes

Indexes are generally located at the end of a taxonomic work. There can be a single index, or multiple indexes following each other. Indexes do not necessarily need to consist of subjects and page numbers; indexes of vernacular names are also

possible. Indexes are often followed by information on the original printer (e.g. "Printed by...etc."); this can be removed too.

You should check whether everything that is at the end of a volume is actually an index, as sometimes addenda are added after an index. Obviously, these should not be removed. Sometimes geographic maps with an index of locations on that map are also present. These should also be kept.

To remove one or more indexes and what follows them, do the following:

- 1) First, you click at the start of the first index, at the beginning of the first word (Figure 46).
- 2) Then you scroll to the end of the document - use the scroll wheel on the mouse, and take care not to click anywhere in the document.
- 3) Now, you press the "Shift"-key on your keyboard and hold it down. Click at the end of the text. All of the text is now selected (Figure 47).
- 4) Press the "Delete"-key on your keyboard to delete the selected text.

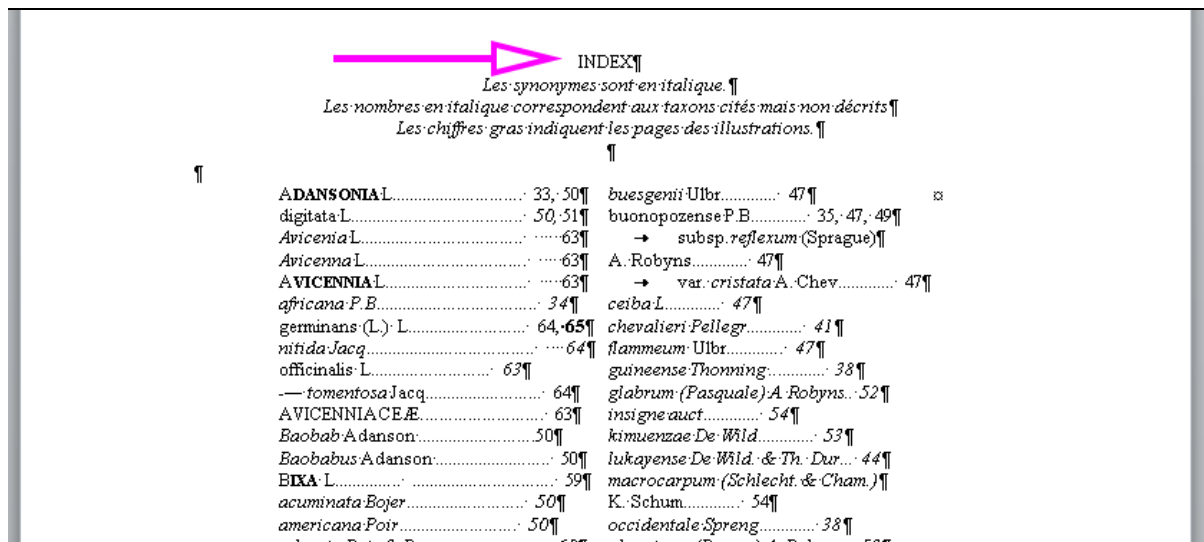


Figure 46: Where to click to start selecting a whole index.



Figure 47: A selection going to the end of the volume.

Should strange things happen with the text position or the number of columns when you delete the indexes, then it is likely there are some strangely placed page, section, or column breaks. In that case, undo your previous deleting action and delete bit by bit. You may need to experiment a little bit to solve such a problem. See also the part below about page, section, and column breaks.

Text formatting that interferes with mark-up

In this section various text formatting problems and their solutions are discussed that can occur in a legacy taxonomic work text.

Interrupted text

Often paragraphs are interrupted in the text of legacy taxonomic works. Usually, the reason for this is that paragraph boundaries were incorrectly recognised during the OCR process. In some cases it is better to treat two separate lines as one line or paragraph to facilitate the semi-automated mark-up process.

Figure 48 shows a few examples at the start of a taxonomic text. At the top of the figure, there is a title listing six family names that is spread over two lines. For the mark-up it is better to have these in a single line. The same applies to the address somewhat lower. In the list of abbreviations some lines have been interrupted too early, which needs to be fixed. Figure 49 shows the corrected lines.

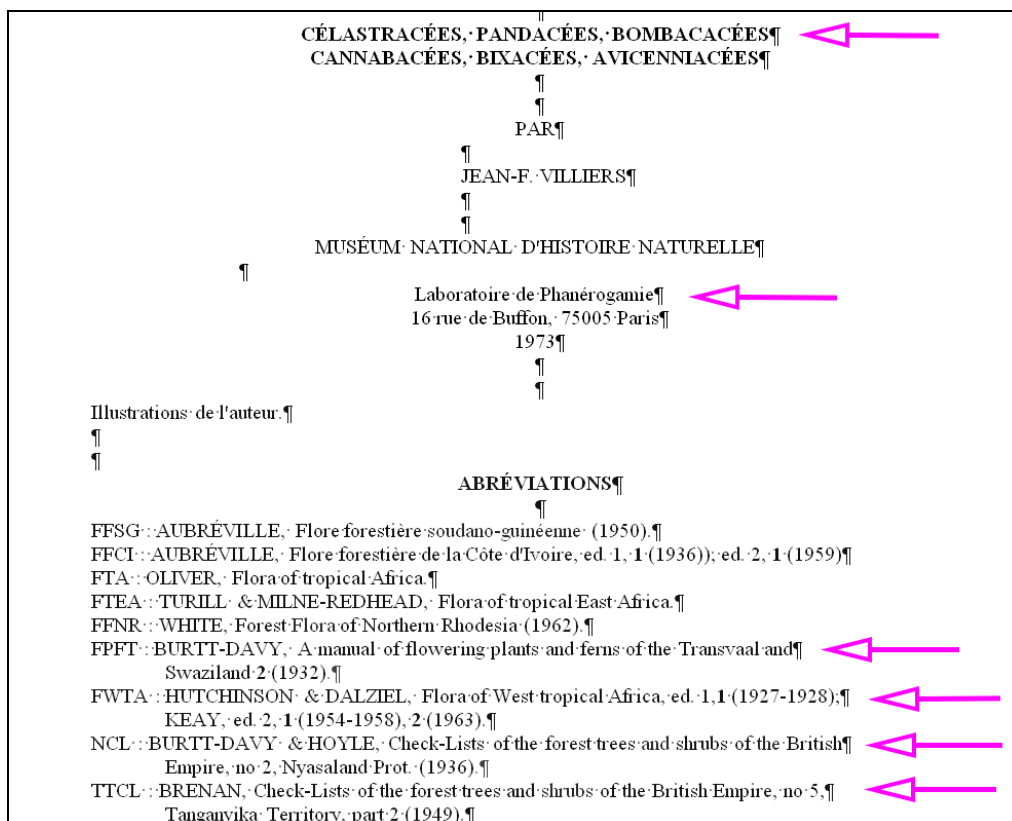


Figure 48: Some examples of interrupted text in a legacy taxonomic work.

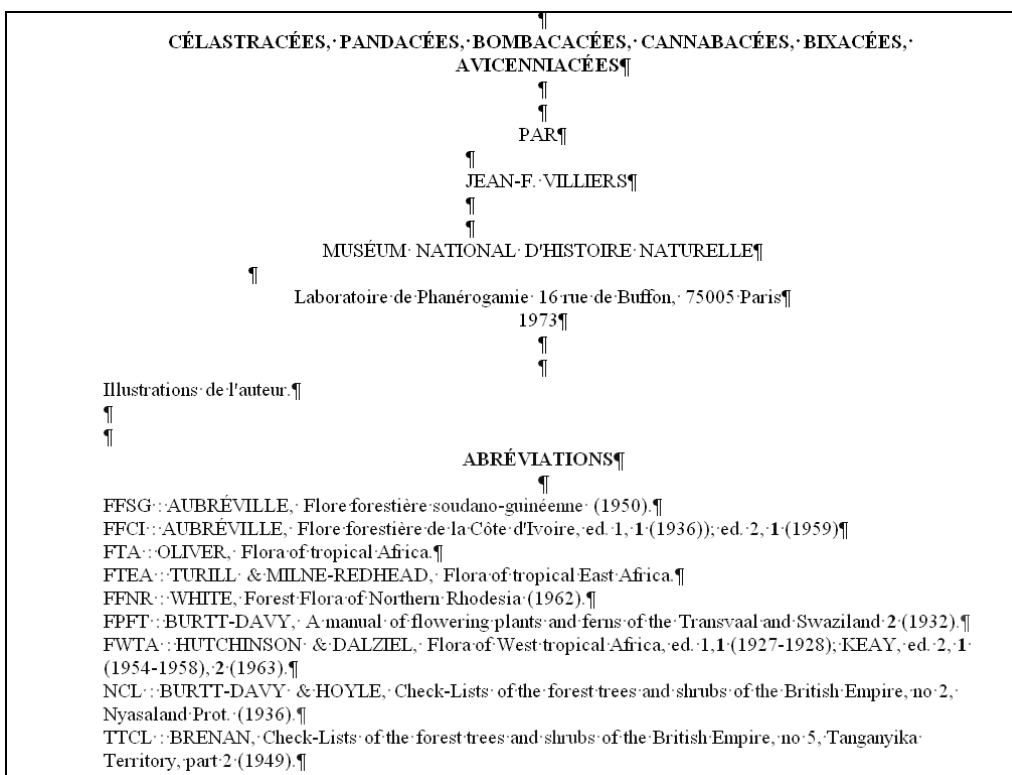


Figure 49: The same examples after correction.

In some cases a figure caption can be present exactly in the middle of a paragraph (Figure 50).

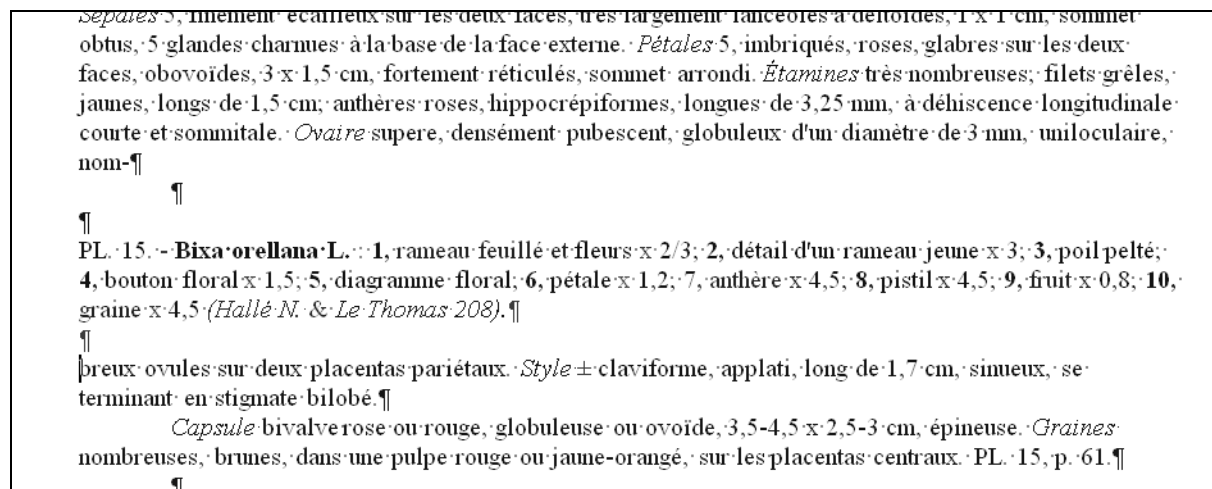


Figure 50: A figure caption interrupts a taxon description.

In such cases, you should do the following:

- 1) First, you select the figure caption and move it to a better place, in this case after the taxon description. Then you select the paragraph marks in between the two paragraph halves (Figure 51).
- 2) You delete the paragraph marks to re-join the two paragraph halves (Figure 52). In this particular case you would also remove the dash (-) and join the two halves of the word.

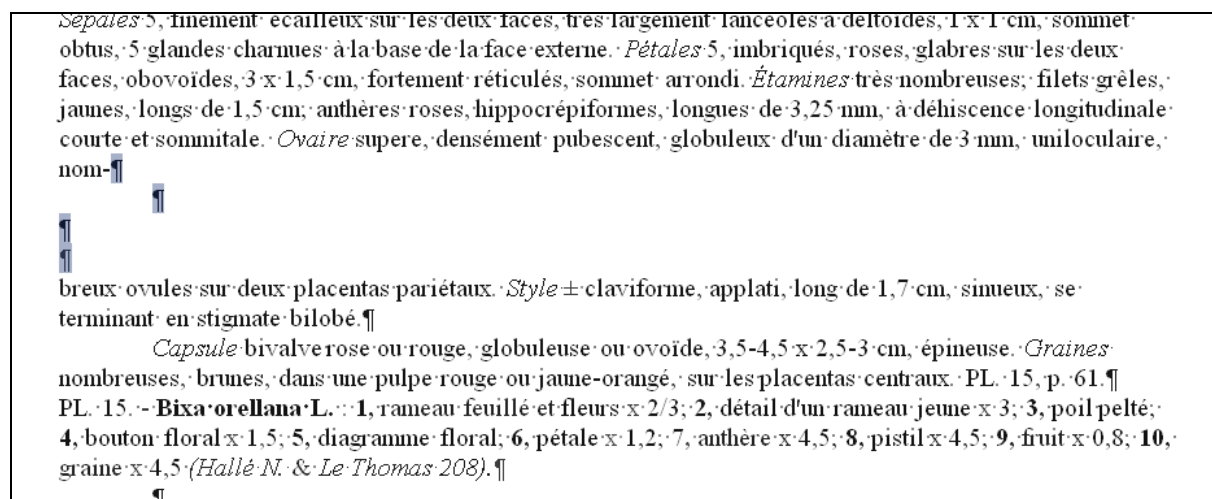


Figure 51: Fixing a figure caption interrupting text: First move the figure caption.

Sépales 5, finement écaillés sur les deux faces, très largement lanceolés à deltoïdes, 1 x 1 cm, sommet obtus, 5 glandes charnues à la base de la face externe. *Pétales* 5, imbriqués, roses, glabres sur les deux faces, obovoïdes, 3 x 1,5 cm, fortement réticulés, sommet arrondi. *Étamines* très nombreuses; filets grêles, jaunes, longs de 1,5 cm; anthères roses, hippocrépiformes, longues de 3,25 mm, à déhiscence longitudinale courte et sommitale. *Ovaire* supère, densément pubescent, globuleux d'un diamètre de 3 mm, uniloculaire, nombreux ovules sur deux placentas pariétaux. *Style* ± claviforme, aplati, long de 1,7 cm, sinueux, se terminant en stigmat bilobé. ¶

Capsule bivalve rose ou rouge, globuleuse ou ovoïde, 3,5-4,5 x 2,5-3 cm, épineuse. *Graines* nombreuses, brunes, dans une pulpe rouge ou jaune-orangé, sur les placentas centraux. PL. 15, p. 61. ¶

PL. 15. ~ **Bixa orellana** L.: 1, rameau feuillé et fleurs x 2/3; 2, détail d'un rameau jeune x 3; 3, poil pelté; 4, bouton floral x 1,5; 5, diagramme floral; 6, pétale x 1,2; 7, anthère x 4,5; 8, pistil x 4,5; 9, fruit x 0,8; 10, graine x 4,5 (Hallé N. & Le Thomas 208). ¶

Figure 52: Fixing a figure caption interrupting text: Re-joining the two halves of the description.

This can also happen with tables. You treat this in an analogous fashion.

Interrupted text more often occurs (and is harder to spot) in keys, nomenclature and literature references than in regular text.

Missing paragraph marks

The opposite of what was described in the previous section is also possible: missing paragraph marks. Figure 53 shows an example in a key. You can simply solve these problems by putting the cursor at the right point and pressing the "Enter"-key once (Figure 54).

2a. Leaves with 1-2(-4) pairs of pinnae, each with (6-)12-25(-29) pairs of leaflets; leaflets 0.6-1.6(-2.3) by 1.5-3.2(-5.5) mm *P. juliflora* b. Leaves with (1-)3-4 pairs of pinnae, each with 6-15 pairs of leaflets; leaflets 2.5-8.3 by 1.4-4 mm *P. pallida* ¶

Figure 53: Missing paragraph mark in key couplet.

2a. Leaves with 1-2(-4) pairs of pinnae, each with (6-)12-25(-29) pairs of leaflets; leaflets 0.6-1.6(-2.3) by 1.5-3.2(-5.5) mm *P. juliflora* ¶

b. Leaves with (1-)3-4 pairs of pinnae, each with 6-15 pairs of leaflets; leaflets 2.5-8.3 by 1.4-4 mm *P. pallida* ¶

Figure 54: After a paragraph mark has been inserted at the right place.

Missing paragraph marks more often occur (and are harder to spot) in keys, nomenclature and literature references than in regular text.

Page, section, and column breaks

You will need to remove all page breaks, section breaks, and column breaks. In some cases you will have to perform some additional work, and in some other cases the break is not immediately obvious. These are all discussed below.

1) Page breaks:

- a. The simplest page break is one where the text is not interrupted by the page break, such as in Figure 55. In this case you can select the page break and delete it.

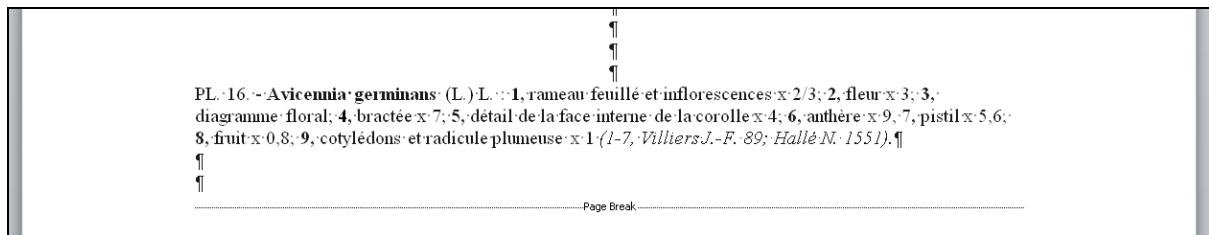


Figure 55: Clean page break.

- b. A somewhat more complex case occurs when the page break happens at a point where the paragraph immediately preceding it is not actually finished (Figure 56). This is similar to the interrupted text examples given earlier, so after deleting the page break you should join the two paragraph halves together.

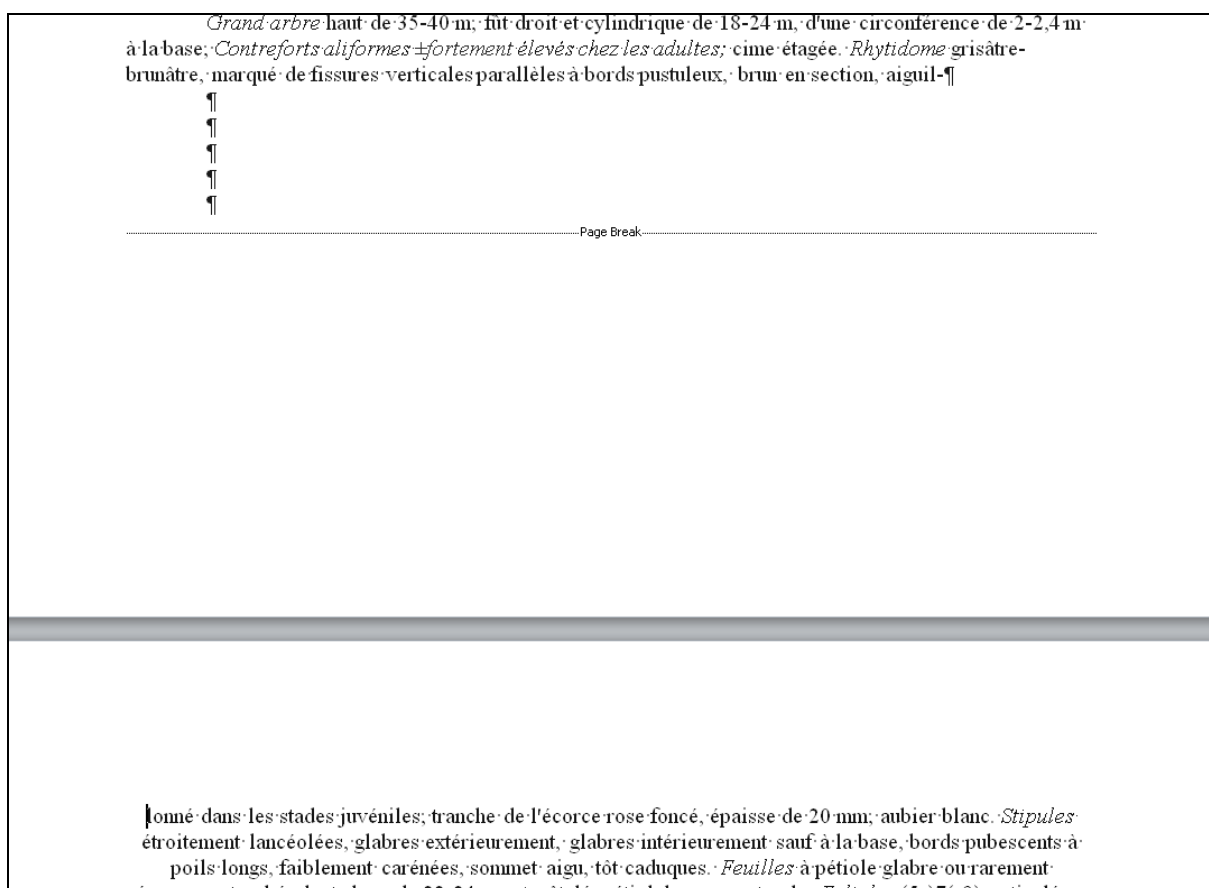


Figure 56: Page break interrupting a paragraph of text.

- c. It is also possible to have a combination of situation b) and the figure caption interrupting the text. You can fix this by first deleting the page

breaks, then moving the figure caption, and then joining the paragraph halves.

2) Section breaks:

- a. The simplest section break is one where you can see there is a section break present (Figure 57). You can just select and delete it.

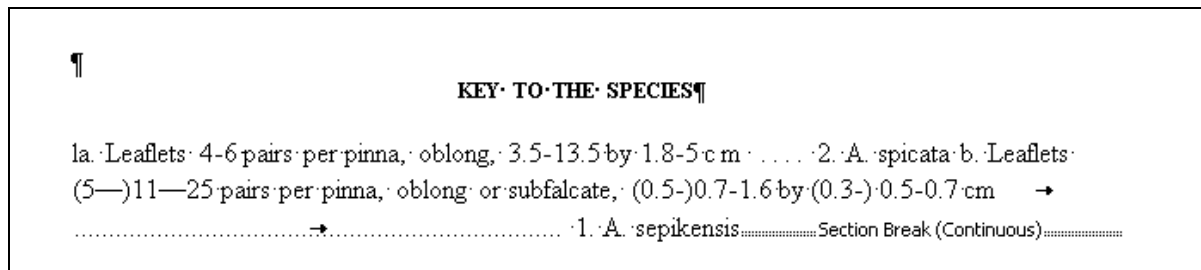


Figure 57: A section break.

- b. Other section breaks are not visible themselves, but you can deduct one is present by looking at the text position within a text column (Figure 58) or the positions of columns amongst each other (Figure 59).

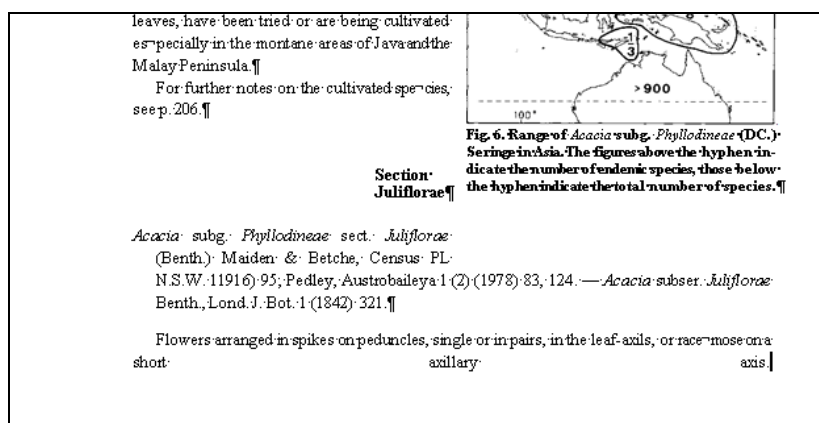


Figure 58: One indication of the presence of a section break.

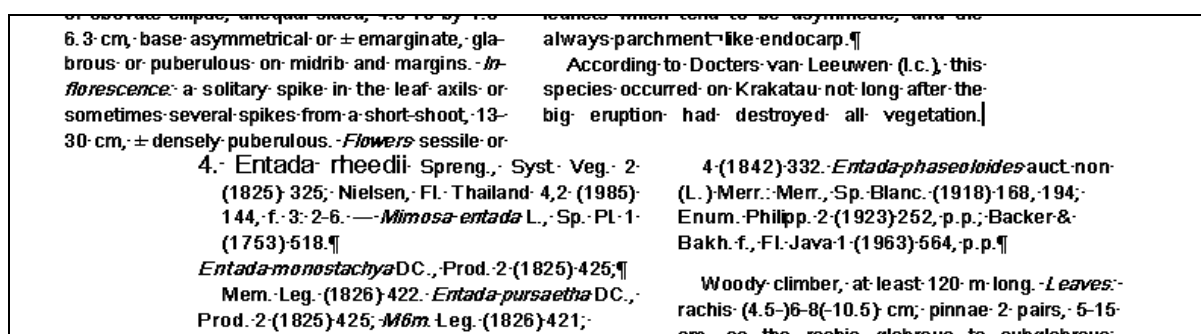


Figure 59: Another indication of the presence of a section break.

- c. These sections breaks can be revealed by putting the cursor at the point where you suspect the break is and pressing the "Enter"-key

once or twice (Figure 60, Figure 61). You can then select them and delete them.

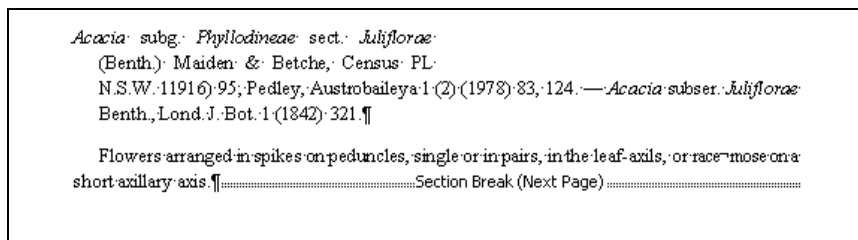


Figure 60: One hidden section break revealed.

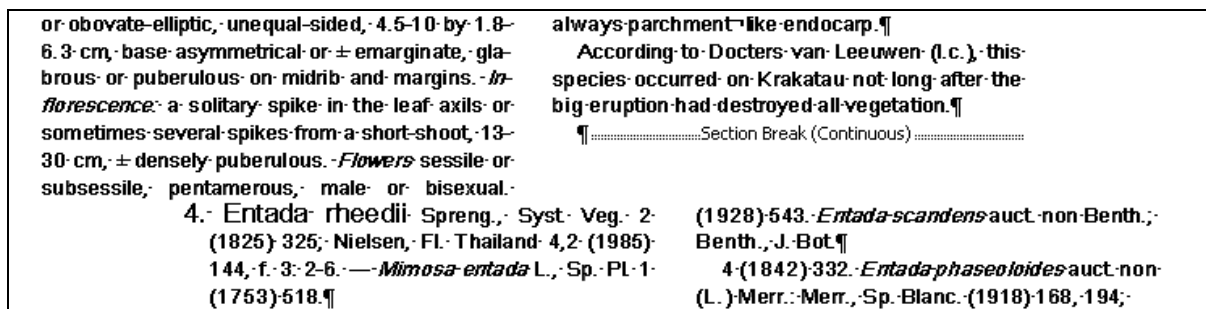


Figure 61: The other hidden section break revealed.

- d. Sometimes it seems section breaks cannot be selected or deleted. In such cases it is likely that there is another page or section break that follows the first one. Select both at once and delete them together.
- 3) Column breaks:
 - a. A column break is shown in Figure 62. You just select these and delete them.

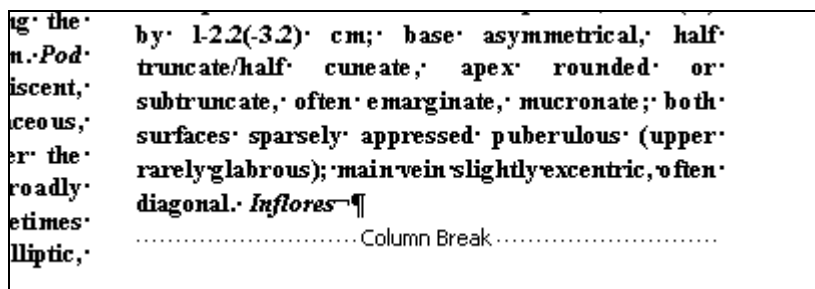


Figure 62: A column break.

Page and section breaks can cause strange effects if you are not careful when you remove them. If this happens, undo your last action using the key combination "Ctrl-Z" and try again.

Tip: If you suspect there is a page, section, or column break present that causes strange things to happen, but you cannot find it at all, you may want to switch the page view from "Print Layout" to "Draft". To do this, go to the "View"-tab and click on the "Draft"-button. To switch back to

"Print Layout", click on the "Print Layout"-button (also in the "View"-tab).

Particularly hard to find page/section/column breaks can also be revealed by removing styles from a paragraph. Instructions on how to do this are given below the Removing styles heading.

Text in text boxes

When text appears to be vertically offset compared to the text column(s) next to it, it is likely said text is actually placed in a text box. Figure 63 shows two examples of this. The circled text is supposed to be one line in the left column that is on the same height as the text in the right column. Instead, it is all over the place. One of the text boxes has been selected in Figure 64.

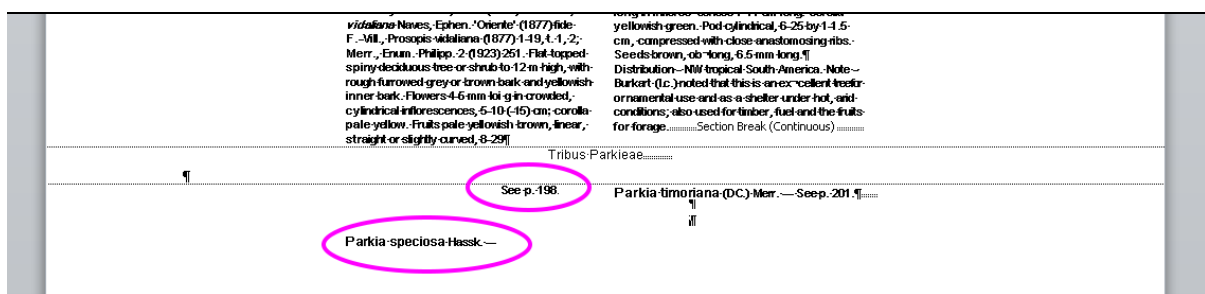


Figure 63: Text that actually is in text boxes.

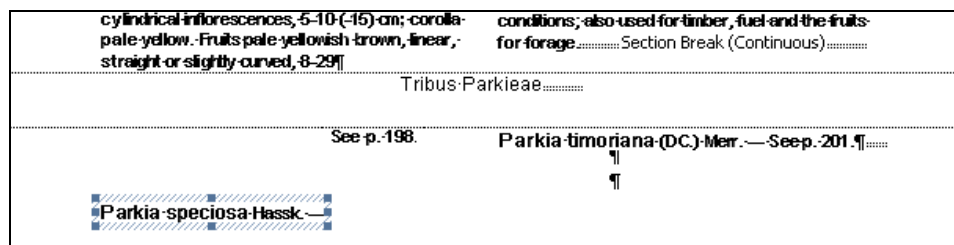


Figure 64: Text in text box, with one text box selected.

Resolving this is similar to moving figure captions out of text boxes, except that due to the large amount of section breaks nearby some care must be exercised.

- 1) First, you have to insert a blank line at the best suited location in the main text (Figure 65, next page). Here the text box text will be pasted.

(continued next page)

pale yellow. Fruits pale yellowish brown, linear, straight or slightly curved, 8-29¶	for forage.....Section Break (Continuous).....
Tribus Parkieae.....	
See p. 198.	¶ Parkia timoriana (DC.) Merr. — See p. 201. ¶.....
Parkia speciosa Hassk. —	¶

Figure 65: Inserting a blank line for the text from the text boxes.

- 2) You click on the text box containing the first part of the text, and select the text (Figure 66).

inner bark. Flowers 4-6 mm l o i g in crowded, cylindrical inflorescences, 5-10 (-15) cm; corolla pale yellow. Fruits pale yellowish brown, linear, straight or slightly curved, 8-29¶	ornamental use and as a shelter under hot, and conditions; also used for timber, fuel and the fruits for forage.....Section Break (Continuous).....
Tribus Parkieae.....	
See p. 198.	¶ Parkia timoriana (DC.) Merr. — See p. 201. ¶.....
Parkia speciosa Hassk. —	¶

Figure 66: Selecting the first part of the text.

- 3) Now you cut the text out of the text box by pressing "Ctrl-X", and paste it at the intended location using "Ctrl-V" (Figure 67).

rough furrowed grey or brown bark and yellowish inner bark. Flowers 4-6 mm l o i g in crowded, cylindrical inflorescences, 5-10 (-15) cm; corolla pale yellow. Fruits pale yellowish brown, linear, straight or slightly curved, 8-29¶	Burkart (l.c.) noted that this is an excellent tree for ornamental use and as a shelter under hot, and conditions; also used for timber, fuel and the fruits for forage.....Section Break (Continuous).....
Tribus Parkieae.....	
See p. 198.	Parkia speciosa Hassk. —¶
	Parkia timoriana (DC.) Merr. — See p. 201. ¶.....
	¶

Figure 67: Pasting the text at the correct location.

- 4) You repeat steps two and three with the second text box (Figure 68).

rough-furrowed grey or brown bark and yellowish inner bark. Flowers 4-6 mm long in crowded, cylindrical inflorescences, 5-10 (-15) cm; corolla pale yellow. Fruits pale yellowish brown, linear, straight or slightly curved, 8-29¶	Burkart (l.c.) noted that this is an excellent tree for ornamental use and as a shelter under hot and conditions; also used for timber, fuel and the fruits for forage. Section Break (Continuous)
Tribus Parkieae.....	
Parkia speciosa Hassk. — See p. 198.¶ Parkia timoriana (DC.) Merr. — See p. 201.¶ ¶ ¶	

Figure 68: The text from the text boxes has been moved to the main text.

- 5) You can now check whether the text boxes have disappeared. If not, delete them.

Footnotes using Word's footnote feature

Generally speaking, footnotes work as follows:

- 1) In the main text you use a small superscript symbol, such as a number or asterisk(s), to indicate there is a footnote - basically a reference to a footnote.
- 2) The footnotes are located at the bottom of the page, preceded by their respective number(s) or asterisk(s).

Microsoft Word has a function to simplify the insertion of footnotes. The problem with footnotes inserted using this function is that when a document using them is converted to plain text (see "Saving the cleaned up text file for processing with Perl scripts"), the footnotes and the symbols referencing them are lost. Obviously this is not the intention.

The solution is to move the footnotes to the main text. Since this will unfortunately remove the numbers or asterisks referencing them, those will have to be retyped. The instructions below explain how:

- 1) First, you have to be able to recognize when footnotes have been inserted using Microsoft Word's footnote function. Figure 69 shows a page from Flora Malesiana with three footnotes at the bottom. The first belongs to the page's title and the two others to the authors. Each reference number (at the top of the page) is encased in a small box with dotted edges. This is always the case with footnotes inserted using Microsoft Word's footnote function.
 - a. If you spot a superscript number that lacks the box around it, the footnote was not inserted using Microsoft Word's footnote function. In that case you should just check whether it is in a footer or textbox. If that is the case, you should cut it out and paste it in the main text.

(continued next page)

■ **MIMOSACEAE (LEGUMINOSAE - MIMOSOIDEAE)**¶¶

(I.C. Nielsen[§], Aarhus, Denmark; H.C. Fortune Hopkins[§], Chatham-Maritime, U.K.)¶¶

Trees, shrubs or lianas, very rarely herbs (*Nephelia* and *Mimosa* p.p.); branches unarmed or armed with stipular thorns (rarely axillary thorns) or scattered prickles on the internodes. Stipules rarely absent, usually caducous. Leaves alternate, usually bipinnate (unipinnate in *Inga*, transformed into phyllodes in *Acacia* subg. *Phyllodineae*), usually provided with extrafloral nectaries on rachis and pinnae. Inflorescences bracteate, simple or compound, racemose; inflorescence units usually consisting of pedunculate glomerules, spikes or spike-like racemes, which are aggregated into axillary or terminal panicles. Pedicels usually short or absent. Flowers actinomorphic, bisexual, unisexual, or rarely neuter, usually small and white, greenish or yellow. Disk, when present, intrastaminal. Stamens few to numerous, free or united into a tube, the latter sometimes united with the corolla tube at the base. Anthers dorsifixed, ± quadrangular in outline, sometimes with a small, caducous gland at the apex. Ovary(-ies) solitary (to several and free), superior, 1-celled; style filiform; stigma small, tubular (infundibular), terminal. Ovules anatropous, parietal. Fruit a pod, dehiscent or indehiscent, sometimes breaking into 1-seeded segments. Seeds usually in two rows from the single placenta, inserted transversely, obliquely or longitudinally, mostly ovate-obicular in outline, often compressed; funicle rarely developed into an aril (*Acacia* p.p., *Pithecellobium*); the testa osseous, coriaceous or chartaceous usually with a ± peripheral furrow, the pleurogram.¶¶

Distribution. — About 60 genera and some 3000 species, mainly in the tropics and the subtropics, but some genera (e.g. *Acacia* and *Albizia*) extending into the warm-temperate zone; in Malesia: 19 genera, of which 15 native, with 1 endemic, viz. *Wallacodendron* in N Celebes and the Philippines. Among the remaining 14 native genera, 5 are pantropical (*Acacia*, *Albizia*, *Entada*, *Nephelia*, *Parkia*), 3 are shared with continental S. Asia and tropical N. Australia (*Adenanthera*, *Archidendron*, *Cathormion*), 2 with Melanesia and the west-Pacific (*Schleinitzia*, *Serianthes*), 2 with Australia (*Pararchidendron*, *Paraserianthes*), 1 with New Caledonia, the Solomon Islands and Australia (*Archidendropsis*), and 1 with India and tropical Africa/Madagascar (*Dichrostachys*). The total number of native and naturalized species is c. 150. Furthermore, an enumeration of c. 45 cultivated species is given at the end of this revision (p. 205). In both Keys to the genera 7 commonly cultivated genera are included.¶¶

In the family *Mimosaceae* tropical Asia and Australia have close affinities, a number of species being common to E. Malesia and tropical (to subtropical) Australia. The links between Asia and Africa are weak, although a few species (*Acacia nilotica*, *Dichrostachys cinerea*, *Entada rheedii*) and a part of the very diversified genus *Calliandra* are common to both continents. Other links between Asia and Africa are *Xylia* from India/Burma to Thailand/Indochina and the rest of the species in Africa/Madagascar and the genera of the *Adenanthera* group, *Adenanthera* being endemic to Asia-Australia and *Tetrapleura* and *Amblygonocarpus* to tropical Africa. The only generic tie between Asia and tropical America is the not yet fully understood *Havardia*: 3 species in mainland Asia, the remaining c. 20 in Central and N. tropical South America (Nielsen 1981).

¶¶

§ — Dedicated to the memory of Dr. Rob Geesink (1944-1992).¶¶

§ — The Danish Natural Science Research Council made this study possible by grants for both travel in Borneo and Java and salary for the first author; a support that hereby gratefully is acknowledged. Initially, Professor C. G. J. van Steenis was very helpful in raising these funds.¶¶

§ — Revision of the genus *Parkia*.¶¶

Figure 69: An example page showing footnotes and their references (at the top of the page).

- 2) You are going to move the footnotes to the main text. Unfortunately, it is not possible to move all of the footnotes at once. Instead, you need to move the text of each footnote separately.
- 3) Start by inserting some blank lines in the main text at the point you want to move the footnotes to. Try to do this as close to the footnote references as possible without interrupting any paragraphs. Type the number of each footnote manually (Figure 70). Do not use Word's numbered list function!

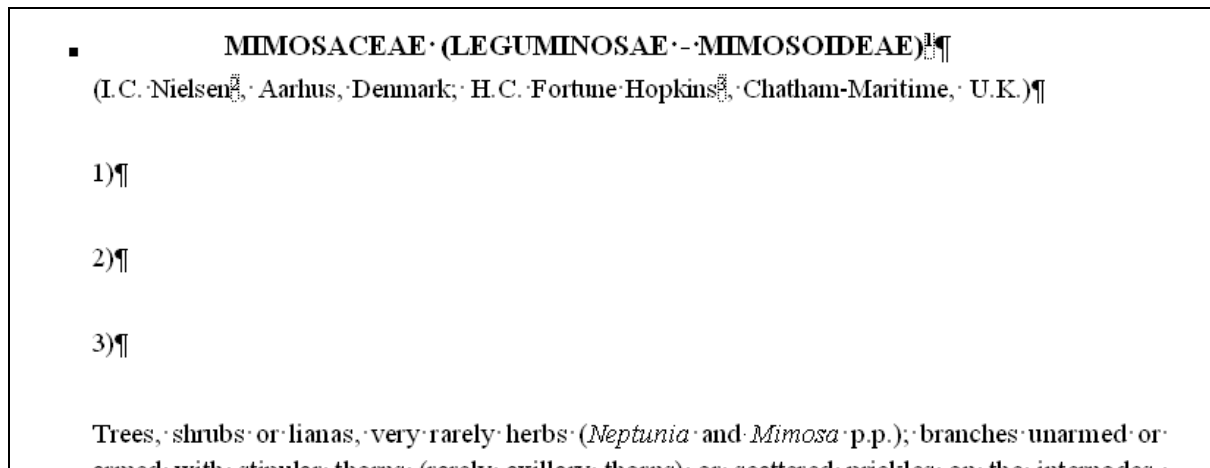


Figure 70: Inserting manually numbered lines for the footnote text.

- 4) Now you go to the footnotes, and select the text of the first footnote (Figure 71).

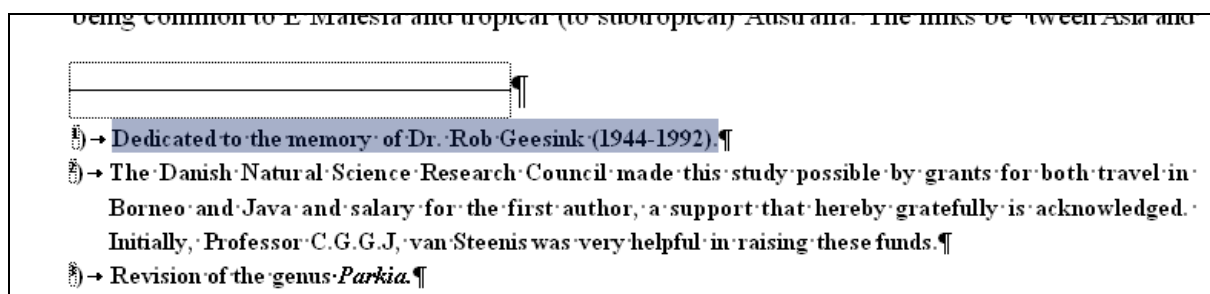


Figure 71: Selecting the first footnote.

- 5) You then press "Ctrl-X" on the keyboard (first press the "Ctrl"-key, keep it down, press the "X"-key, release both keys) to cut out the text. Then you go to the first line you inserted at step three, and you paste the text there using "Ctrl-V" (Figure 72).
- 6) You repeat steps four and five for the other two footnotes.

(continued next page)

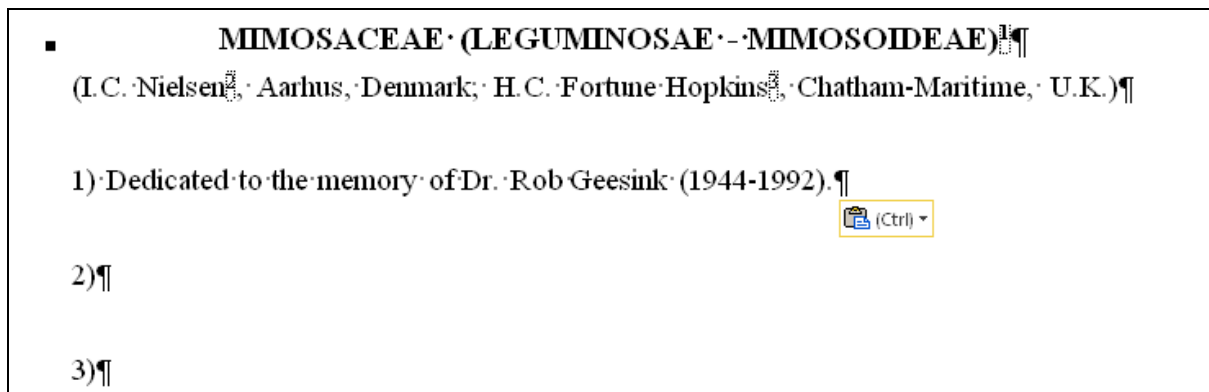


Figure 72: Pasting the first footnote in the main text.

- 7) Now you click next to each footnote reference number, and type in the number by hand (Figure 73).

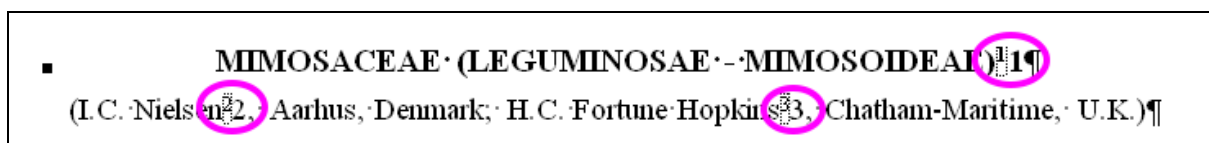


Figure 73: Manually typed footnote references.

- 8) You then select all of the original footnote reference numbers and delete them (Figure 74).

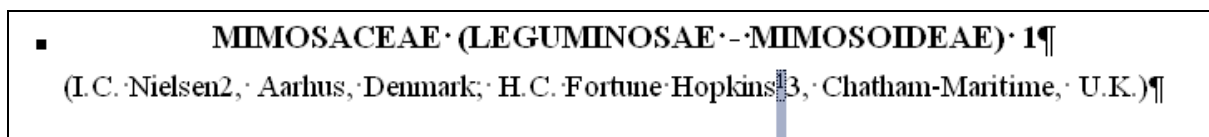


Figure 74: Selecting and removing the original footnote reference numbers.

- 9) Normally this should also remove the left-over bits of the footnotes at the bottom of the page. If not, clean up.

Manual line breaks

Manual line breaks sometimes are present in legacy taxonomic work texts (Figure 75) and have to be removed.

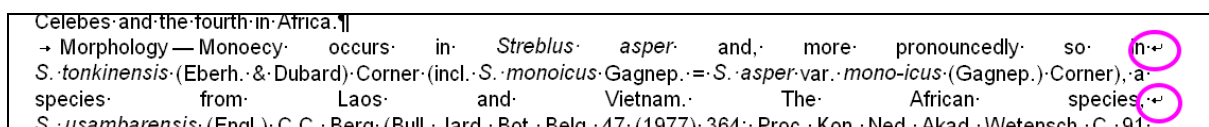


Figure 75: Two manual line breaks.

You can remove them by selecting them and pressing the "Delete"-key.

However be careful, as they sometimes also occur in text that is supposed to be in two columns but was recognized as being in one column.

Column issues

Rarely, text that was supposed to be part of two columns is recognized as being part of one column during the OCR process (Figure 76). The only way to fix this is to very carefully reassemble both columns using cut and paste, with the original paper work used as a reference.

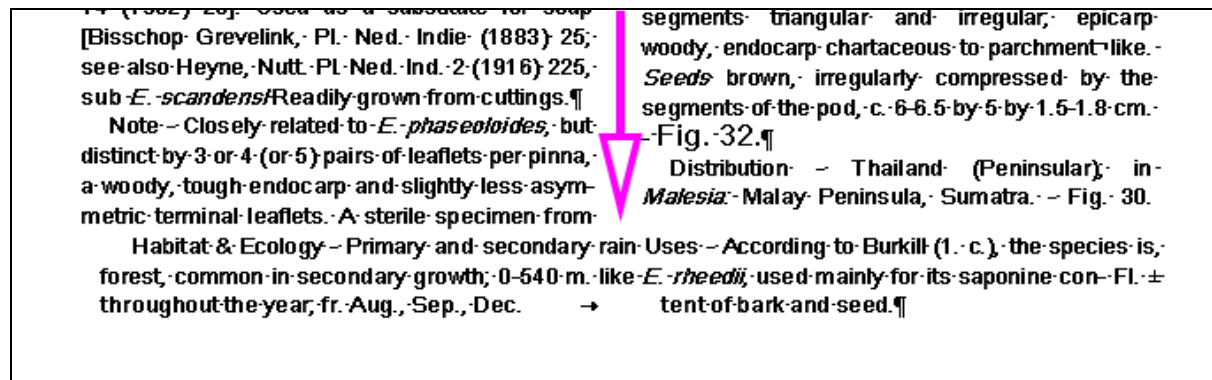


Figure 76: Text that is supposed to be in two columns has erroneously been recognized as being in a single column. The arrow indicates where the whitespace between the columns is supposed to be.

Tables

Tables have to be prepared for mark-up by ensuring that all the information that has to go in one table cell is on a single line. Below are instructions on how to do this:

- 1) Figure 77 shows a table as you may find it in a legacy taxonomic work that has been OCR'ed. As you can see, the contents of the two last cells of the fifth row are split over two lines (Figure 78).

	<i>B. minutiflora</i>	→	<i>B. fruticosa</i> ¶
Taille de la plante.....→.....	non précisée	→	arbuste de 1,5 m¶
Inflorescence.....→.....	15 cm	→	7-10 cm¶
Extérieur du calice.....→.....	glabre, cilié au bord	→	très finement pubérulent¶
Filets des étamines.....→.....	très courts, aussi larges que l'anthère	→	très courts, rétrécis sous l'anthère¶
Longueur des tépales.....→.....	2/3 du périanthe	→	moitié du périanthe.¶

Figure 77: A table as it can be found in a legacy taxonomic work, in this case Flore du Gabon.

	<i>B. minutiflora</i>	→	<i>B. fruticosa</i> ¶
Taille de la plante.....→.....	non précisée	→	arbuste de 1,5 m¶
Inflorescence.....→.....	15 cm	→	7-10 cm¶
Extérieur du calice.....→.....	glabre, cilié au bord	→	très finement pubérulent¶
Filets des étamines.....→.....	très courts, aussi larges que l'anthere	→	très courts, rétrécis sous l'anthere¶
Longueur des tépales.....→.....	2/3 du périanthe	→	moitié du périanthe.¶

Figure 78: Contents of two cells split over two lines.

- 2) First, you will ensure that the text of each cell is sequential. For this, select the text as shown in Figure 79 and move it as shown in Figure 80. Repeat this for the other cell and remove the remains of the second line (Figure 81).

	<i>B. minutiflora</i>	→	<i>B. fruticosa</i> ¶
Taille de la plante.....→.....	non précisée	→	arbuste de 1,5 m¶
Inflorescence.....→.....	15 cm	→	7-10 cm¶
Extérieur du calice.....→.....	glabre, cilié au bord	→	très finement pubérulent¶
Filets des étamines.....→.....	très courts, aussi larges que l'anthere	→	très courts, rétrécis sous l'anthere¶
Longueur des tépales.....→.....	2/3 du périanthe	→	moitié du périanthe.¶

Figure 79: The text that is going to be moved is selected.

	<i>B. minutiflora</i>	→	<i>B. fruticosa</i> ¶
Taille de la plante.....→.....	non précisée	→	arbuste de 1,5 m¶
Inflorescence.....→.....	15 cm	→	7-10 cm¶
Extérieur du calice.....→.....	glabre, cilié au bord	→	très finement pubérulent¶
Filets des étamines.....→.....	très courts, aussi larges que l'anthere	→	très courts, rétrécis sous l'anthere¶
Longueur des tépales.....→.....	2/3 du périanthe	→	moitié du périanthe.¶

Figure 80: After moving the text.

	<i>B. minutiflora</i>	→	<i>B. fruticosa</i> ¶
Taille de la plante.....→.....	non précisée	→	arbuste de 1,5 m¶
Inflorescence.....→.....	15 cm	→	7-10 cm¶
Extérieur du calice.....→.....	glabre, cilié au bord	→	très finement pubérulent¶
Filets des étamines.....→.....	très courts, aussi larges que l'anthere	→	très courts, rétrécis sous l'anthere¶
Longueur des tépales.....→.....	2/3 du périanthe	→	moitié du périanthe.¶

Figure 81: After moving the other piece of text and removing the second line.

- 3) Now ensure that the text for each cell is on a separate line, as shown in Figure 82. Excess whitespace may be removed.

<i>B. minutiflora</i>
<i>B. fruticosa</i>
Taille de la plante
non précisée
arbuste de 1,5 m
Inflorescence
15 cm
7-10 cm
Extérieur du calice
glabre, cilié au bord
très finement pubérulent
Filets des étamines
très courts, aussi larges que l'anthère
très courts, rétrécis sous l'anthère
Longueur des tépales
2/3 du périanthe
moitié du périanthe

Figure 82: Table prepared for mark-up.

- 4) It is also possible that the table has been clearly recognized as a table by the OCR process and that the table therefore uses Microsoft Word's table formatting (Figure 83).

	<i>B. minutiflora</i>	<i>B. fruticosa</i>
Taille de la plante	non précisée	arbuste de 1,5 m
Inflorescence	15 cm	7-10 cm
Extérieur du calice	glabre, cilié au bord	très finement pubérulent
Filets des étamines	très courts, aussi larges que l'anthère	très courts, rétrécis sous l'anthère
Longueur des tépales	2/3 du périanthe	moitié du périanthe

Figure 83: A table formatted using Microsoft Word's table formatting options

- 5) This is usually the case when a table has clearly drawn borders in the original printed document. In this case, you have to insert as many blank lines as you need table cells below the table (Figure 84).

(continued next page)

Lists using Microsoft Word's list function

In bulleted and numbered lists using Microsoft Word's list function, the bullets or numbers will be lost when the document is saved as a plain text file. Obviously this is not the intention. This means that you will have to keep an eye out for these, and fix them whenever they occur. The instructions below explain how to do this:

- 1) You can recognize lists that use Microsoft Word's list function by each list item being a bullet or number followed by ostensibly a tab, except that the bullet or number cannot actually be edited when you click on it (Figure 87).

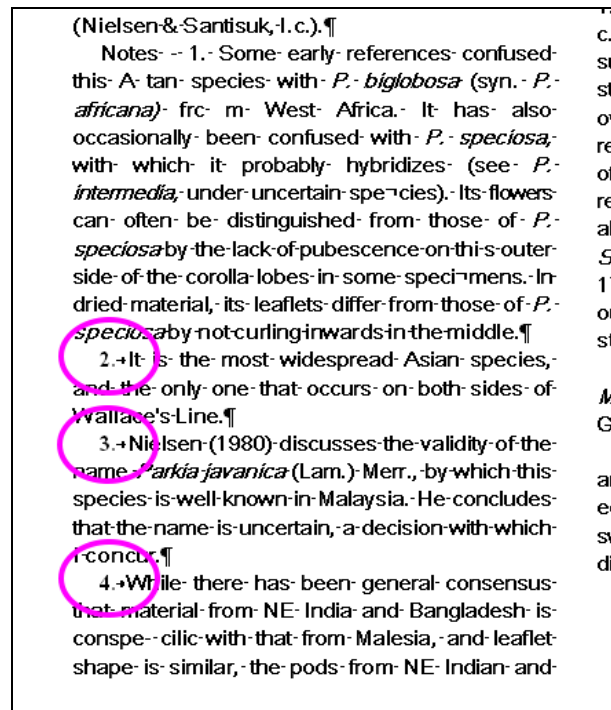


Figure 87: A numbered list using Microsoft Word's list function.

- 2) To fix this, you start by removing the original bullet or number (Figure 88). Start with the last list item in numbered lists, as removing a number early in the list will cause the others to be renumbered.

(continued next page)

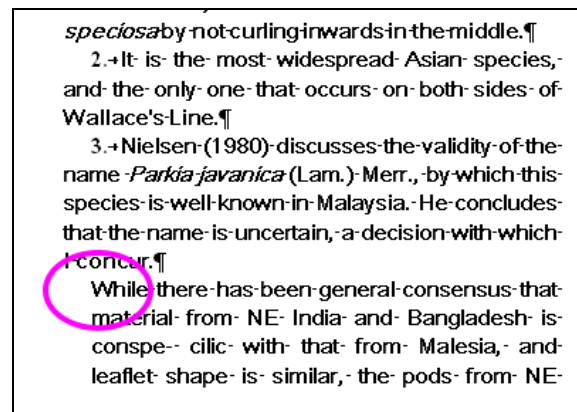


Figure 88: Removing the last number in a numbered list using Microsoft Word's list function.

- 3) Then you manually type the list number the list item should have (Figure 92). For numbers, use the number followed by a dot and space. For bulleted lists, you can use a short dash (-) followed by a space.

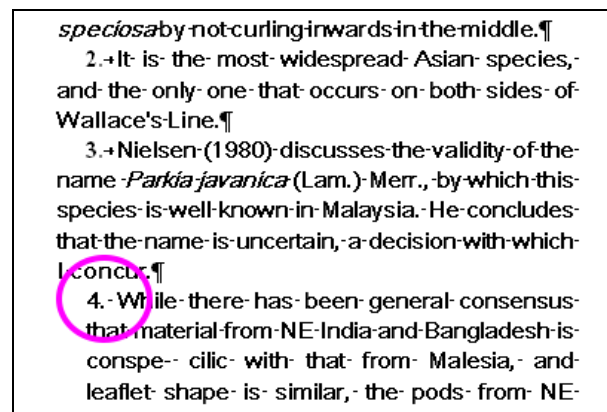


Figure 89: Manually type the list item number.

- 4) You then repeat step two and three for the other list items that need to be changed (Figure 90).

(continued next page)

(Nielsen & Santusuk, l.c.).¶

Notes -- 1. Some early references confused this Asian species with *P. biglobosa* (syn. *P. africana*) from West Africa. It has also occasionally been confused with *P. speciosa*, with which it probably hybridizes (see *P. intermedia*, under uncertain species). Its flowers can often be distinguished from those of *P. speciosa* by the lack of pubescence on the outside of the corolla lobes in some specimens. In dried material, its leaflets differ from those of *P. speciosa* by not curling inwards in the middle.¶

2. It is the most widespread Asian species, and the only one that occurs on both sides of Wallace's Line.¶

3. Nielsen (1980) discusses the validity of the name *Parkia javanica* (Lam.) Merr., by which this species is well known in Malaysia. He concludes that the name is uncertain, a decision with which I concur.¶

4. While there has been general consensus that material from NE India and Bangladesh is conspecific with that from Malaysia, and leaflet shape is similar, the pods from NE

Figure 90: Corrected list with manually typed numbers.

Text formatted as a numbered list using Microsoft Word's list function is often present in both actual lists and list-like text like keys.

Text wrongly recognized as a list

In some cases, text can be wrongly recognized as a list during the OCR process when it clearly is none, as shown in Figure 91. The only solution for this is to retype the affected parts of the text.

flat, membranaceous to coriaceous, glabrous or with sparse short pubescence margins, 1.5–2.3 by 0.5–0.8 cm, rounded to the stipe. *Seeds*¶

3. → 8 → per pod, brown, ovoid-orbicular, compressed¶

4. → 4.1 → by 3.4–3.5 mm, areole open towards the hilum.¶

Distribution – A large part of Australia; in *Malesia*: Philippines (Luzon), Les (Flores, Timor), New Guinea (Papua New Guinea).¶

Figure 91: Text wrongly recognized as a list during the OCR process.

What can be left in?

Excessive whitespace, multiple empty lines and punctuation errors like ;; can be left in. Frequently occurring OCR errors can also be left in.

Because spotting and removing all the more rarely occurring OCR errors is very tedious and hard, these errors are only removed when spotted. Do not try to proofread hundreds of pages of taxonomic text.

Renumbering indented keys

Many older taxonomic works contain indented keys; keys in which the hierarchy of the key is determined by the amount of whitespace preceding each line.

Unfortunately, this whitespace often does not survive the OCR process unscathed, seriously complicating the creation of a Perl script to automate the mark-up of such keys. It was found that it was less time-consuming to convert indented keys to linked keys than to try to fix the whitespace issues. Therefore you will have to convert any indented key to a linked key in the text you are preparing.

- 1) The simplest of indented keys is shown in Figure 92. In this case there is only one couplet consisting of two question leads, making it very simple to renumber the question leads. **1** becomes **1a**, **1'** becomes **b** (Figure 93). You do the renumbering by selecting the number and typing in the new number.

CLÉ DES GENRES¶	
1. Feuilles subopposées; rameaux à 4 ailes; cyme multiflore lâche, de grande taille; stipules trifides; capsule à 4-5 valves; graine fortement saillante hors de l'arille.....	1. <i>EUONYMUS</i> ¶
1'. Feuilles alternes; rameaux à 3 ailes; cyme pauciflore de courte taille; stipules simples; capsule à 2 valves; graine totalement recouverte par l'arille.....	2. <i>MAYTENUS</i> ¶

Figure 92: The simplest indented key possible.

CLÉ DES GENRES¶	
1a. Feuilles subopposées; rameaux à 4 ailes; cyme multiflore lâche, de grande taille; stipules trifides; capsule à 4-5 valves; graine fortement saillante hors de l'arille.....	1. <i>EUONYMUS</i> ¶
b. Feuilles alternes; rameaux à 3 ailes; cyme pauciflore de courte taille; stipules simples; capsule à 2 valves; graine totalement recouverte par l'arille.....	2. <i>MAYTENUS</i> ¶

Figure 93: The simplest key possible after conversion to a linked key.

- 2) A key with two couplets is shown in Figure 94. In this case you will want to renumber as follows: **1** becomes **1a**, **1'** becomes **b**, **2** becomes **2a**, **2'** becomes **b**.
- 3) However, at that point you do not have a linked key yet. The b-lead of couplet 1 should point to couplet 2. So you add a 2 at the end of the b-lead. Make sure there is a space in front of the '2' (Figure 95).

CLÉ DES ESPÈCES PAR LES FEUILLES¶	
1. Folioles pétiolulées.....	3. <i>P. aquatica</i> ¶
1'. Folioles sessiles ou subsessiles.¶	
→ 2. Face inférieure du limbe pubérulente ou lépidote.....	1. <i>P. glabra</i> ¶
→ 2'. Face inférieure du limbe glabre.....	2. <i>P. sessilis</i> ¶

Figure 94: An indented key with two couplets.

CLÉ DES ESPÈCES PAR LES FEUILLES	
1a. Folioles pétiolulées.....	3. <i>P. aquatica</i>
b. Folioles sessiles ou subsessile.....2	
...2a. Face inférieure du limbe pubérulente ou lépidote.....	1. <i>P. glabra</i>
...b. Face inférieure du limbe glabre.....	2. <i>P. sessilis</i>

Figure 95: The same key after conversion. Note the circled 2 that links the b-lead of the first couplet to the second couplet.

- 4) In indented keys that consist of more than two couplets, question leads will have to be moved around besides having to renumber and link them. Figure 96 shows a short indented key with four couplets.

1. D'APRÈS LES ÉCHANTILLONS ♂	
1. Rameaux ± fortement pubescents.	
...2. Pétales glabres intérieurement.....	1. <i>M. klainei</i>
...2'. Pétales pubérulents intérieurement.	
...3. Pistillode pubescent.....	2. <i>M. pierlotiana</i>
...3'. Pistillode glabre.....	3. <i>M. puberula</i>
1'. Rameaux glabres ou très fortement glabrescents.	
...4. Pétales pubescents ou pubérulents intérieurement.....	4. <i>M. haumaniana</i>
...4'. Pétales glabres intérieurement.....	5. <i>M. camerunensis</i>

Figure 96: An indented key with four couplets.

- 5) It is best to work couplet by couplet and question lead by question lead. You first renumber lead 1 to 1a. Then you add a 2 to the end of the 1a-lead to link it to couplet 2 (Figure 97).

1. D'APRÈS LES ÉCHANTILLONS ♂	
1a. Rameaux ± fortement pubescents. 2	
...2. Pétales glabres intérieurement.....	1. <i>M. klainei</i>
...2'. Pétales pubérulents intérieurement.	
...3. Pistillode pubescent.....	2. <i>M. pierlotiana</i>
...3'. Pistillode glabre.....	3. <i>M. puberula</i>
1'. Rameaux glabres ou très fortement glabrescents.	
...4. Pétales pubescents ou pubérulents intérieurement.....	4. <i>M. haumaniana</i>
...4'. Pétales glabres intérieurement.....	5. <i>M. camerunensis</i>

Figure 97: Step by step conversion of an indented key with multiple couplets to a linked key: Renumbering the first lead of the first couplet and linking it to the second couplet.

- 6) Then you go to lead 1'. You renumber this to b, and then link it to couplet 4 by adding a 4 at the end of the line (Figure 98).

1. D'APRÈS LES ÉCHANTILLONS ♂	
1a. Rameaux ± fortement pubescents. 2	
2. Pétales glabres intérieurement	1. <i>M. klainei</i>
2'. Pétales pubérulents intérieurement.	
3. Pistillode pubescent.	2. <i>M. pierlotiana</i>
3'. Pistillode glabre.	3. <i>M. puberula</i>
b. Rameaux glabres ou très fortement glabrescents. 4	
4. Pétales pubescents ou pubérulents intérieurement	4. <i>M. haumaniana</i>
4'. Pétales glabres intérieurement	5. <i>M. camerunensis</i>

Figure 98: Step by step conversion of an indented key with multiple couplets to a linked key: Renumbering the second lead of the first couplet and linking it to the third couplet.

7) Now you move the b-lead of the first couplet to its proper location (Figure 99).

1. D'APRÈS LES ÉCHANTILLONS ♂	
1a. Rameaux ± fortement pubescents. 2	
b. Rameaux glabres ou très fortement glabrescents. 4	
2. Pétales glabres intérieurement	1. <i>M. klainei</i>
2'. Pétales pubérulents intérieurement.	
3. Pistillode pubescent.	2. <i>M. pierlotiana</i>
3'. Pistillode glabre.	3. <i>M. puberula</i>
4. Pétales pubescents ou pubérulents intérieurement	4. <i>M. haumaniana</i>
4'. Pétales glabres intérieurement	5. <i>M. camerunensis</i>

Figure 99: Step by step conversion of an indented key with multiple couplets to a linked key: Moving the second lead of the first couplet to its proper location.

8) You can now repeat the steps above for the second (Figure 100), third and fourth couplets (Figure 101).

1. D'APRÈS LES ÉCHANTILLONS ♂	
1a. Rameaux ± fortement pubescents. 2	
b. Rameaux glabres ou très fortement glabrescents. 4	
2a. Pétales glabres intérieurement	1. <i>M. klainei</i>
b. Pétales pubérulents intérieurement. 3	
3. Pistillode pubescent.	2. <i>M. pierlotiana</i>
3'. Pistillode glabre.	3. <i>M. puberula</i>
4. Pétales pubescents ou pubérulents intérieurement	4. <i>M. haumaniana</i>
4'. Pétales glabres intérieurement	5. <i>M. camerunensis</i>

Figure 100: Step by step conversion of an indented key with multiple couplets to a linked key: Taking care of the second couplet.

1. D'APRÈS LES ÉCHANTILLONS ♂ ¶	
1a. Rameaux ± fortement pubescents. 2¶	
b. Rameaux glabres ou très fortement glabrescents. 4¶	
2a. Pétales glabres intérieurement	1. <i>M. klainei</i> ¶
b. Pétales pubérulents intérieurement. 3¶	
3a. Pistillode pubescent.	2. <i>M. pierlotiana</i> ¶
b. Pistillode glabre.	3. <i>M. puberula</i> ¶
4a. Pétales pubescents ou pubérulents intérieurement	4. <i>M. haumaniana</i> ¶
b. Pétales glabres intérieurement	5. <i>M. camerunensis</i> ¶

Figure 101: Step by step conversion of an indented key with multiple couplets to a linked key: Converted key.

- 9) In indented keys with more than four couplets you can obviously follow the same approach. However, it is very important to do this work in a systematic and attentioned manner, as especially in larger indented keys it is possible to become very confused. It is recommended that you keep a paper copy of the taxonomic work you are working on at hand as a reference.

Some additional notes:

- The number or letter at the start of each lead should always include a dot. So "1a.", "b." etc.
- You will rarely have to move leads around when converting indented keys of less than three couplets.
- If the keys contain leads with more than one apostrophe, change the letters used for the leads appropriately. So if you have the following leads: **1**, **1'**, **1''**, **1'''** you renumber them as follows: **1a**, **b**, **c**, **d**.
- **b**, **c**, **d** etc. are not preceded by a number. Only the **a** leads are.

If your document only contains linked keys of the format shown in Figure 101 you do not need to do anything.

Removing styles

Before you save the text in a file format that is suitable for processing with Perl scripts, you may want to remove specific styles from the text. Do the following to do this:

- 1) First, you need to select all the text. Put the mouse cursor anywhere in the text. Press down the "Ctrl"-key on your keyboard, keep it down, and press the "A"-key. Release the keys. All text should be selected.
- 2) Now press down the "Ctrl"-key, hold it down, and press the "Q"-key to remove all styles. Release the keys. All styles should be removed.

You can also remove styles from only part of the text, as it may be used to reveal very hard to find page, section, or column breaks. In that case, you only select that

part of the text from which you want to remove styles and then press "Ctrl-Q" as described above in step two.

Tip: Should the procedures described above for some reason fail to give the desired results (e.g. the line height does not change), you can then go to the "Other styles"-tab, and click on the "Normal" style.

Processing order within the document

It may seem logical that you simply start the clean-up process at the beginning of the text document. Unfortunately, this is not the case. The best place to start the clean-up process is at the end of the text document. This is due to issues with Microsoft Word's ability to deal with format changes; or rather, its inability to do so properly in large documents and its tendency to do very strange things at times when removing certain types of formatting.

So you start at the end (last page) of the document and work your way towards the beginning, going from the bottom of each page to its top (Figure 102).

Note that you can fix some very minor problems without taking this approach, but once the problems involved are on anything more than a very local scale (changing a single character) Word gets in the way.

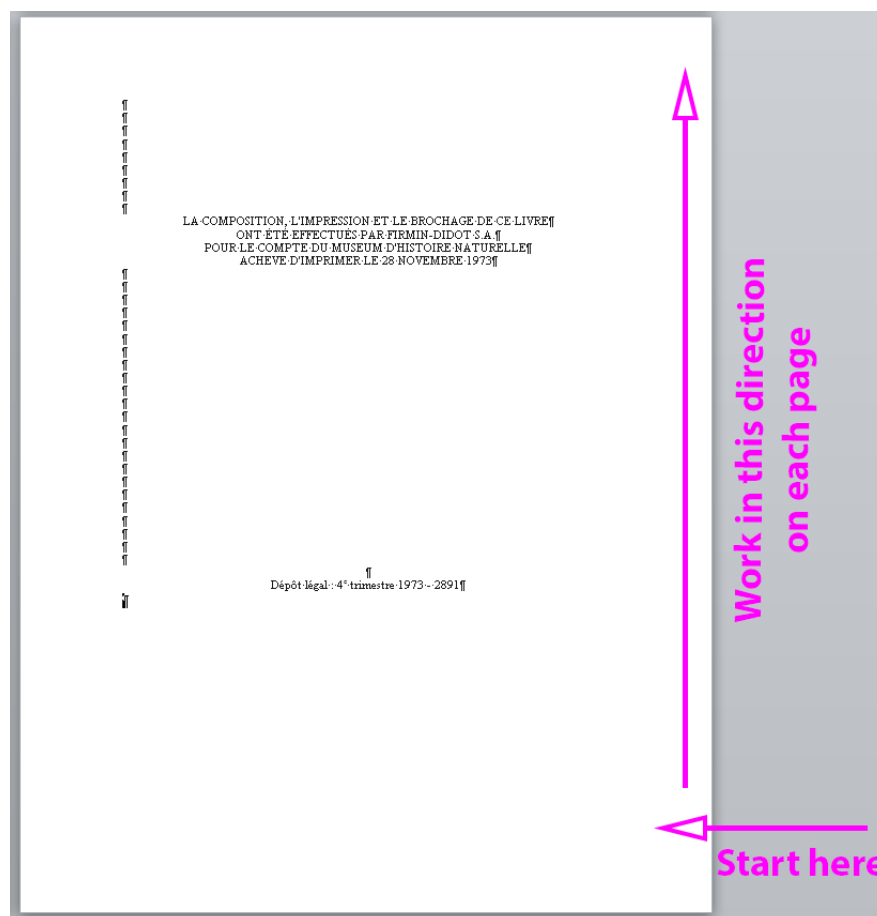


Figure 102: Where to start working on a document (at its end!) and in which direction (from bottom to top).

When working through a page from bottom to top, it is best that you start by looking for text in text boxes and first move that, followed by removing eventual graphics, removing page/column/section breaks, and fixing the left-over issues.

Preparing files for combined mass-processing

One way of speeding up the mark-up process is to combine multiple taxonomic texts in one file, so that they can all be processed together during semi-automated mark-up. To do this, simply paste several cleaned up volumes of a taxonomic work end-to-end into a single document and save as indicated below.

For reasons of practicality during script development/adjustment and proofreading, it is recommended that such combined files are no longer than 1,000-1,500 pages total.

Please do not combine multiple text files prior to clean up; it is possible that Word will do very strange things that will make clean-up much more complicated.

Saving the cleaned up text file for processing with Perl scripts

- 1) To save the cleaned up text file for processing with Perl scripts, go to the "File"-tab and click on "Save As".
- 2) In the "Save As"-window, click on the drop down box next to "Save as type:" and choose "Plain Text (*.txt)" (Figure 103). Then click on the "Save"-button.

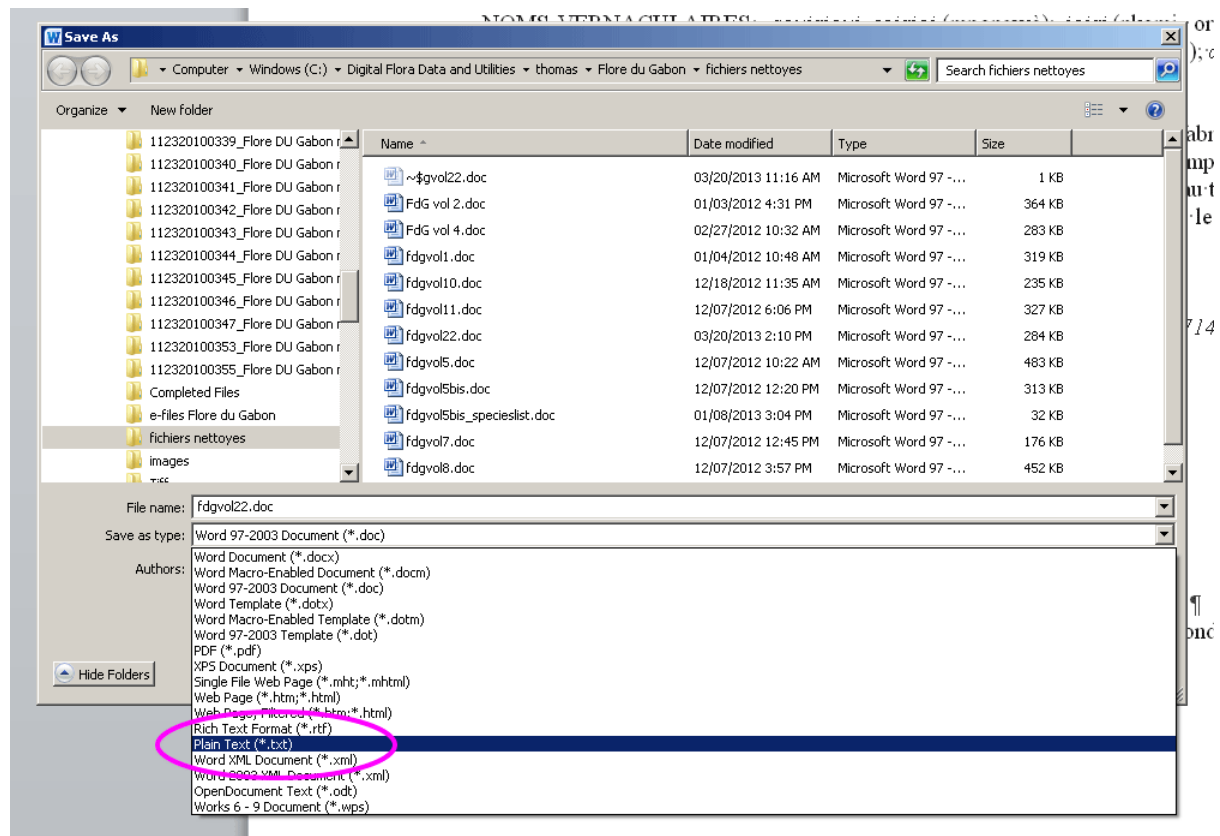


Figure 103: Saving the cleaned text as a text file.

- 3) A "File Conversion"-window appears asking you to select the text encoding (Figure 104). Click on "Other encoding:" and select "Unicode (UTF-8)" from the list. Click on the "OK"-button. The file is now saved.

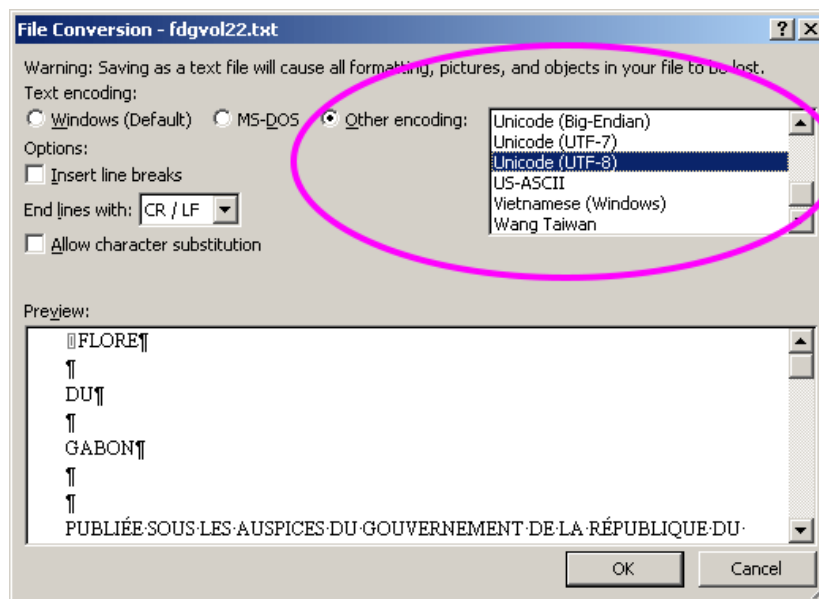


Figure 104: Selecting the proper text encoding.

Why save as Unicode? Unicode is a standardized text encoding that supports many different languages and symbols, including ones that are present in taxonomic works. Other text encodings may not support these symbols, causing that information to be lost.

Appendix I: Perl scripts to further prepare the text

A set of two Perl scripts are used to further prepare the text of the legacy taxonomic work for semi-automated processing. Please refer to **script use.doc** for extensive instructions on how to run Perl scripts. Obviously you need to have Perl installed to make this work.

Running the clean-up script

As was mentioned before, you did not have to remove all kinds of punctuation problems from the text, nor did you have to remove excessive whitespace. This is possible because most of such problems can easily be automated for with a single script that only takes a short moment to run. Short instructions are given below (refer to **script use.doc** for better instructions):

- 1) Use the clean-up script that your script developer has provided you with for that particular taxonomic work.
- 2) The previously saved plain text file is the input file for the clean-up script and should be in the same folder as the script.
- 3) You run the clean-up script from the Command Prompt window by typing
perl [clean-up scriptname.plx] [input_file] [output_file]
(obviously with the proper file names filled in)

Figure 105 shows a sample of a file before a clean-up script was run, while Figure 106 shows the same sample after clean-up with the script.

```

(1-)2(-3) ovules par loge, dressés ou pendants.
58 Capsules loculicides ou drupes, plurispermes en général, parfois moncspermes. Graines arillées ou non, en général albuminées. Embry
59 Cette famille groupe environ 800 espèces dans une soixantaine de genres répartis dans les deux hémisphères. Au Gabon, 2 genres sont
60
61
62
63
64
65
66
67
68
69
70
71 CLÉ DES GENRES
72
73 1a. Feuilles subopposées; rameaux à 4 ailes; cyme multiflore lâche, de grande taille; stipules trifides; capsule à 4-5 valves; gra
74 1. EUONYMUS
75
76 1b. Feuilles alternes; rameaux à 3 ailes; cyme pauciflore de courte taille; stipules simples; capsule à 2 valves; graine totalement
77 . . . . . 2. MAYTENUS
78
79
80 1. EUONYMUS Linné
81
82 -- Sp. Pl. : 197 (1753), « Evonymus » corr. Gen. Pl. ed. 5 : 91 (1754).
83 - Vyenomus PRESL, Abh. Böhm. Ges. Wiss. 3 : 462 (1845).
84 - Melanocarya TURCZANINOW, Bull. Soc. Imp. Nat. Moscou 31 (1) : 453 (1858).
85 - Pragmatessara PIERRE, Fl. For. Cochinch. : tab. 309 (1894).
86 - Pragmatropa PIERRE, I.c. (1894).
87
88 Arbustes dressés ou lianescents, ou arbres à rameaux anguleux ou cylindriques. Feuilles en général opposées. Stipules caduques.
89 Inflorescences en cymes axillaires multiflores ou rarement fleurs solitaires. Fleurs ♂ ou parfois ♀ ou * par avortement, actinomor
90 forme très variable. Étamines en même nombre que les pièces périanthaires, soudées par le filet au disque; anthères introrses à déh
91 disque, 5-loculaire, en général à 2 (-4-10-12) ovules par loge, dressés ou pendants.
92 Capsules à 2-5 valves, déhiscences à (1-3-)4-5 loges, ± charnues à coriaces. Graines 1-2 par loge, entourées d'un arille charnu ou
93
94 ESPÈCE-TYPE : Euonymus europæus Linné.
95

```

Figure 105: Legacy taxonomic work plain text file before running the clean-up script.

26	CELASTRACEÆ
27	(2 genres, 3 espèces)
28	Arbres ou arbustes dressés, parfois lianes, inermes ou épineux. Appareil végétatif et florifère montrant dans certains cas des fils de
29	alternes ou opposées, stipulées, pétiolées ou subsessiles, rarement feuilles verticillées ou absentes.
30	Inflorescences en cymes, plus rarement en grappes ou en panicules, parfois fleurs solitaires ou fasciculées, axillaires ou terminales,
31	tétra- ou pentamères. Sépales contortés, libres ou soudés, parfois absents. Pétales libres ou soudés à la base, contortés. Disque char
32	pétales, alternipétales, insérées sur le disque, setransformant en staminodes ou disparaissant dans les fleurs +; anthères à 2 loges à
33	(1-)2(-3) ovules par loge, dressés ou pendants.
34	Capsules loculicides ou drupes, plurispermes en général, parfois monospermes. Graines arillées ou non, en général albuminées. Embryon
35	Cette famille groupe environ 800 espèces dans une soixantaine de genres répartis dans les deux hémisphères. Au Gabon, 2 genres sont re
36	CLÉ DES GENRES
37	1a. Feuilles subopposées; rameaux à 4 ailes; cyme multiflore lâche, de grande taille; stipules trifides; capsule à 4-5 valves; graine
38	b. Feuilles alternes; rameaux à 3 ailes; cyme pauciflore de courte taille; stipules simples; capsule à 2 valves; graine totalement rec
39	1. EUONYMUS Linné
40	Sp. Pl.: 197 (1753), « Evonymus » corr. Gen. Pl. ed. 5: 91 (1754).
41	- Vytenomus PRESL, Abh. Böhm. Ges. Wiss. 3: 462 (1845).
42	- Melanocarya TURCZANINOW, Bull. Soc. Imp. Nat. Moscou 31 (1): 453 (1858).
43	- Pragmatessara PIERRE, Fl. For. Cochinch.: tab. 309 (1894).
44	- Pragmatropa PIERRE, I.c. (1894).
45	Arbustes dressés ou lianescents, ou arbres à rameaux anguleux ou cylindriques. Feuilles en général opposées. Stipules caduques.
46	Inflorescences en cymes axillaires multiflores ou rarement fleurs solitaires. Fleurs ♂ ou parfois ♀ ou + par avortement, actinomorpe
47	forme très variable. Étamines en même nombre que les pièces périnthaires, soudées par le filet au disque; anthères introrsées à déhisc
48	disque, 5-loculaire, en général à 2 (-4-10-12) ovules par loge, dressés ou pendants.
49	Capsules à 2-5 valves, déhiscences à (1-3-)4-5 loges, ± charnues à coriaces. Graines 1-2 par loge, entourées d'un arille charnu ou mem
50	ESPÈCE-TYPE: Euonymus europæus Linné.
51	Ce genre tropical et tempéré, essentiellement asiatique, comprend environ 220 espèces. Une seule est connue au Gabon.
52	Euonymus congolensis R. Wilczek
53	Bull. Jard. Bot. Etat Bruxelles 29: 183 (1959); Fl. Congo 9: 126, tab. 14 (1960).
54	Liane; tige à 4 ailes, brune, glabre; nombreuses lenticelles claires. Large bourgeon légèrement supra-axillaire, courttement pubérulent
55	courttement pubérulent; pétiole glabre dessus et très courttement pubérulent dessous, concave ou plat. Limbe glabre sur les deux faces
56	X 4,3-2,2 cm, base arrondie ou atténuée, bords latéraux nettement et finement dentés, sommet longuement acuminé-aigu. Nervure médiane
57	alternes, ascendantes, très arquées, s'anastomosant à 2 mm environ du bord du limbe, saillantes dessus et très fortement saillantes de
58	Inflorescences cymeuses axillaires, à ramifications subopposées multiflores; axe très courttement pubérulent; bractéoles lancéolées par
59	pentaédriques au sommet. Bouton floral globuleux, d'un diamètre de 1 à 2 mm; préfloraison tordue. Fleurs ♂ blanches. Sépales largemen
60	mm, sommet obtus. Pétales glabres sur les deux faces, oblongs, ± asymétriques, 3-3,25 X 1,5 mm, sommet obtus, dressés dans la fleur ép
61	membraneux à 5 dents aiguës épinétales, haut de 0,5-1,5 mm; filets libres, grêles, longs de 1 mm; anthères tôt caduques, basifixes, el

Figure 106: Legacy taxonomic work plain text file at the same point after running clean up script.

Running the OCR error fixing script

The other script that you will run at this point is a script that fixes OCR errors that commonly occur in your taxonomic work. Instructions:

- 1) Use the OCR error fixing script that your script developer has provided you with for that particular taxonomic work.
- 2) The output file of the previous step is used as the input file for the OCR error fixing script.
- 3) You run the script from the Command Prompt window by typing
perl [OCRfix scriptname.plx] [input_file] [output_file]
(obviously with the proper file names filled in)

The final manual preparation step: Separating taxa

The final step before you start with the semi-automated mark-up process is to insert blank lines between the taxa. The Perl scripts are not intelligent enough to be able to detect on their own where a new taxon starts, so you have to indicate this for them by inserting a blank line at those points.

The file you have to work in is the output file of the OCR error fixing script. You will do this in Notepad++.

The actual task is very simple:

- 1) First, you start Notepad++. You go to the "File"-menu, choose "Open", and open the output file from the OCR error fixing script (Figure 107). The file in question should be in the same folder as the scripts.

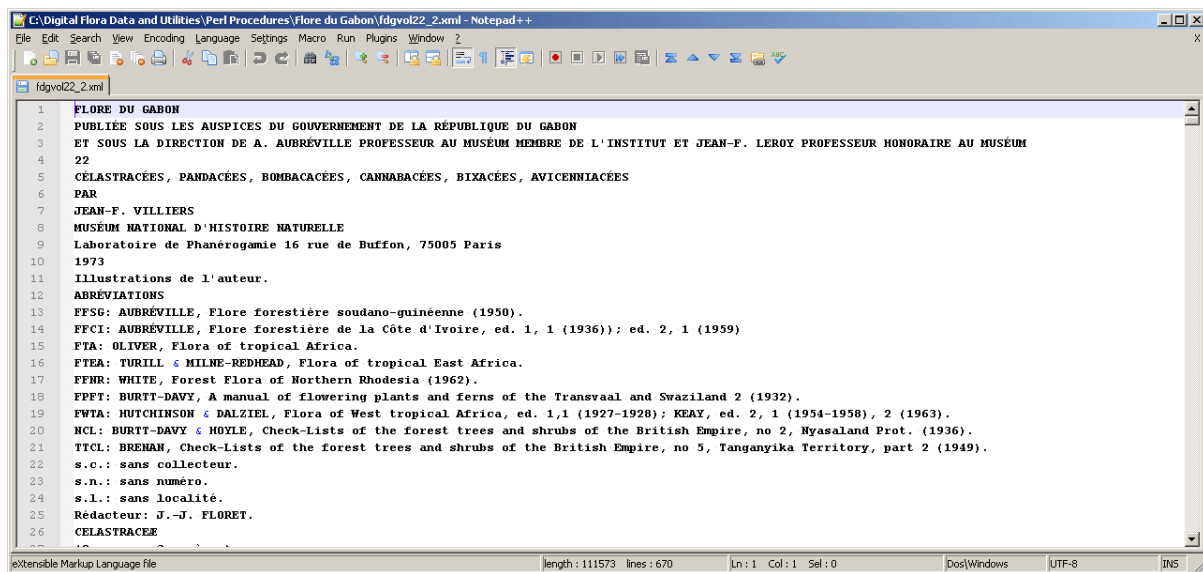


Figure 107: Notepad++ with a taxonomic text file open.

- 2) Now you scroll through the file, and anytime you encounter a new taxon you insert a single blank line. You do this by clicking in front of each new taxon and pressing the "Enter"-key on your keyboard once.
- 3) After a while, you should have something similar to what is shown in Figure 108.

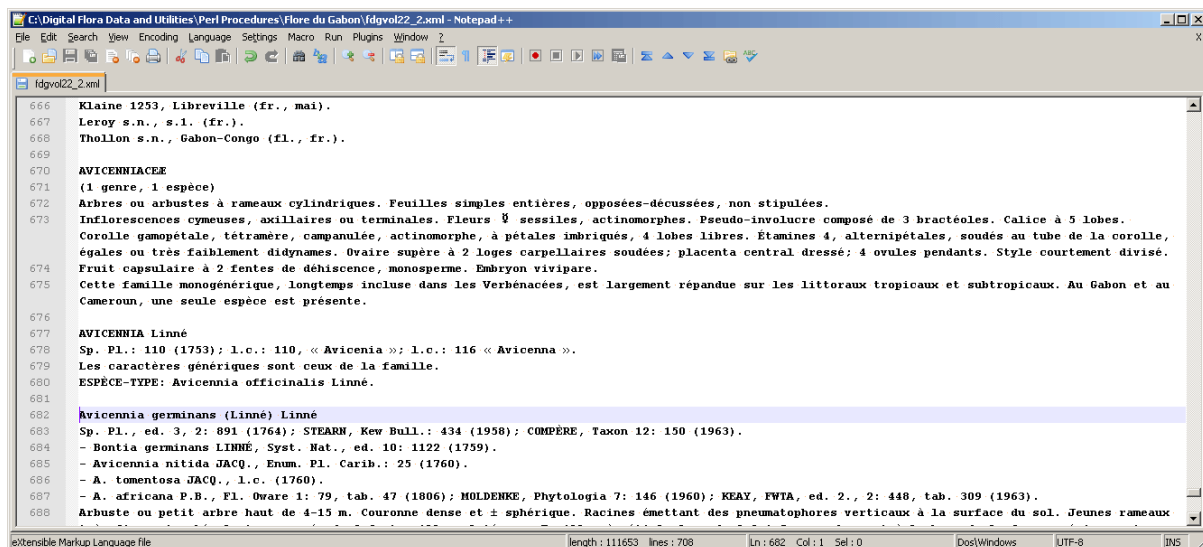


Figure 108: Blank lines between taxa.

Important notes:

- Taxa can have any rank. So if the first taxon you encounter is at family rank, you insert a blank line before it. You then scroll down, and encounter a genus. Again, a blank line before it. Species? The same thing. Etc.
- Do not insert blank lines in the nomenclature that follows a taxon name!
- It is useful to keep a paper copy of the taxonomic work at hand for reference.

- Keys always belong to the taxon that directly precedes them.
- Sometimes taxa are mentioned in a key and then appear to be absent from the rest of the work. This is often the case with varieties or subspecies. In such cases it is useful to check whether their parent species perhaps includes them in a paragraph on taxonomy. If so, copy the name and create a separate taxon for them. If not, ignore.

When you are done with this, check whether there is a blank line at the beginning and/or end of the file you have just edited. If so, delete it.

Do not forget to save the file now and then (and when you are done).

Now you can continue with the actual semi-automated mark-up process. See **script use.doc** for information.

Appendix II: Additional text preparation prior to script running for Flora of the Guianas

The Naturalis flora Flora of the Guianas features some specific types of contents that require special preparation for smoother semi-automated mark-up. This section shortly describes how to handle that contents.

Initial file clean-up

Clean up the file as described in the rest of this document, except do not remove the following two types of indexes:

- Numerical list of accepted taxa (see Figure 109 for an example)
- Collections studied (see Figure 110 for an example)

Do clean up any issues in those two types of indexes, though.

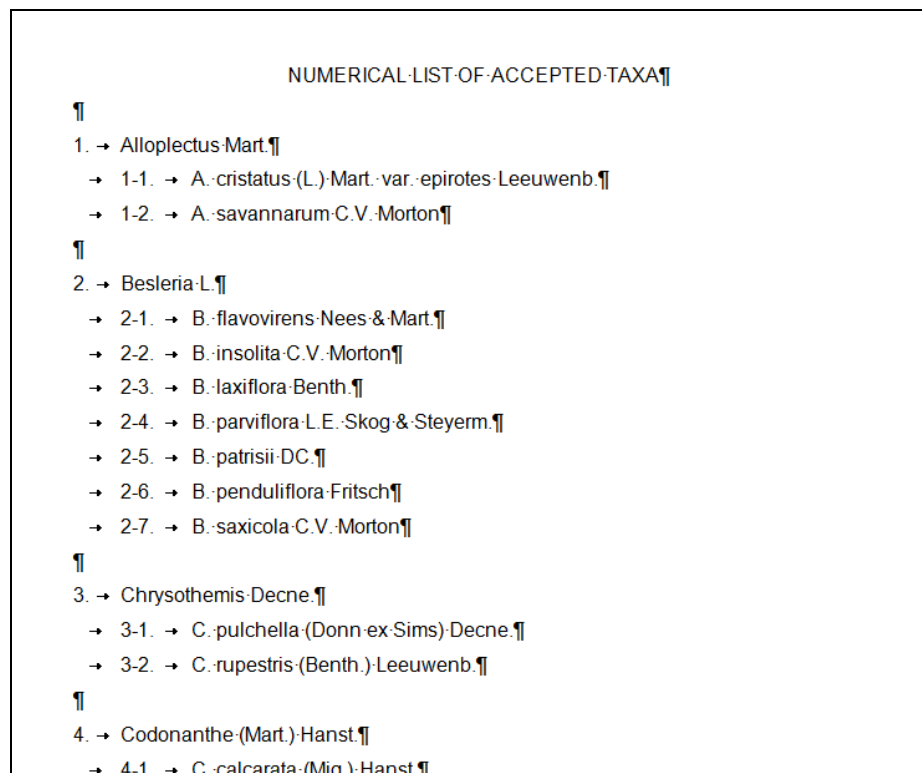


Figure 109: An example of a "Numerical list of accepted taxa"-index in Flora of the Guianas, from Series A volume 26.

COLLECTIONS STUDIED	
(Numbers in bold represent types)	
GUYANA	3); 638 (17-5); 645 (4-1); 725 (17-5); 750 (4-1); 856 (16-7); 1072 (1-2); 1138 (12-1); 1322 (4-2); 1439 (16- 4); 1495, 1515 (4-1); 1790 (19-2); 1820 (3-2); 1928 (17-7); 1929 (9-2); 1959 (4-2); 1973, 2008 (4-1); 2196 (9-2); 2440 (3-1); 2452 (19-1); 2520 (17-5); 2525 (2-3); 2596, 2653 (4-1); 2844 (9-2); 2876 (16-9); 2931, 2932 (2-3); 2961 (16-9); 3056, 3454 (4-2); 3646 (9-2); 3667, 3907 (4-1); 4112 (16-4); 4181 (17-7); 4213 (4-1); 4214 (10-3); 4258 (2-3); 4260 (10- 3); 4369 (9-2); 4392 (16-11); 4408 (4-2); 4454, 4616 (9-2); 4619, 4693 (3-2); 4770 (4-2); 4908 (3-2); 5157
Abraham, A.A., 340 (16-9); 345 (2-7)	
Altson, R.A., 311 (2-7); 321 (16-4); 357 (20-1); 371 (16-2); 422 (16-9)	
Andel, T. van, <i>et al.</i> , 1075 (16-7); 1739 (4-2); 1954 (4-2)	
Anonymous, s.n. (17-7)	
Appun, C.F., s.n. (16-9); 66 (3-2); 2125 (19-1); 5181 (19-1)	
Archer, W.A., 2321 (16-7); 2432 (17-7)	
Atkinson, D.J., 33 (16-4); 77 (17-5)	
Bailey, I.W., 110 (4-1); 162 (16-9); 181 (4-1)	
Bartlett, A.W., 8562 (9-4); 8743 (2-2)	
Beckett, J.E., s.n. (17-7)	

Figure 110: An example of a "Collections studied"-index in *Flora of the Guianas*, from Series A volume 26.

Go to the next step after saving the cleaned up file as Unicode-encoded text.

Preparation of "Collections studied"-index for easier mark-up

As shown in Figure 110, the "Collections studied"-index consists of a list of collections, usually sorted by the area they were collected in. Sometimes there are multiple such lists, one for each family described in a given volume. Each list consists of entries starting with a collector's name (with or without initials). Following this is a semicolon-separated list consisting of the collection number and, between brackets, a numerical code (taxon number) matching the numerical codes found in the "Numerical list of accepted taxa"-index for that given volume (Figure 109).

In this step, you will replace the numerical codes in the "Collections studied"-index with the corresponding taxon names from the "Numerical list of accepted taxa"-index.

Because these numerical codes are repeated whenever there are multiple families per volume (so the same codes are used for different taxa), you will use the following procedure in Notepad++:

- 1) Open the Unicode-encoded (cleaned up) text file in Notepad++.
- 2) Find the "Numerical list of accepted taxa"-index, select it in its entirety, and cut it out of the text file, using the shortcut "Ctrl-X".
 - a. Create a new text file using "Ctrl-N".
 - b. Paste the "Numerical list of accepted taxa"-index into this file. Save the file with a meaningful name.

- c. Now expand all the genus abbreviations found in the “Numerical list of accepted taxa”-index into the full genus name (Figure 111). You will have to do this manually using copy and paste because the same genus abbreviation can be used for multiple genera.
- d. Be sure to save the file once done.

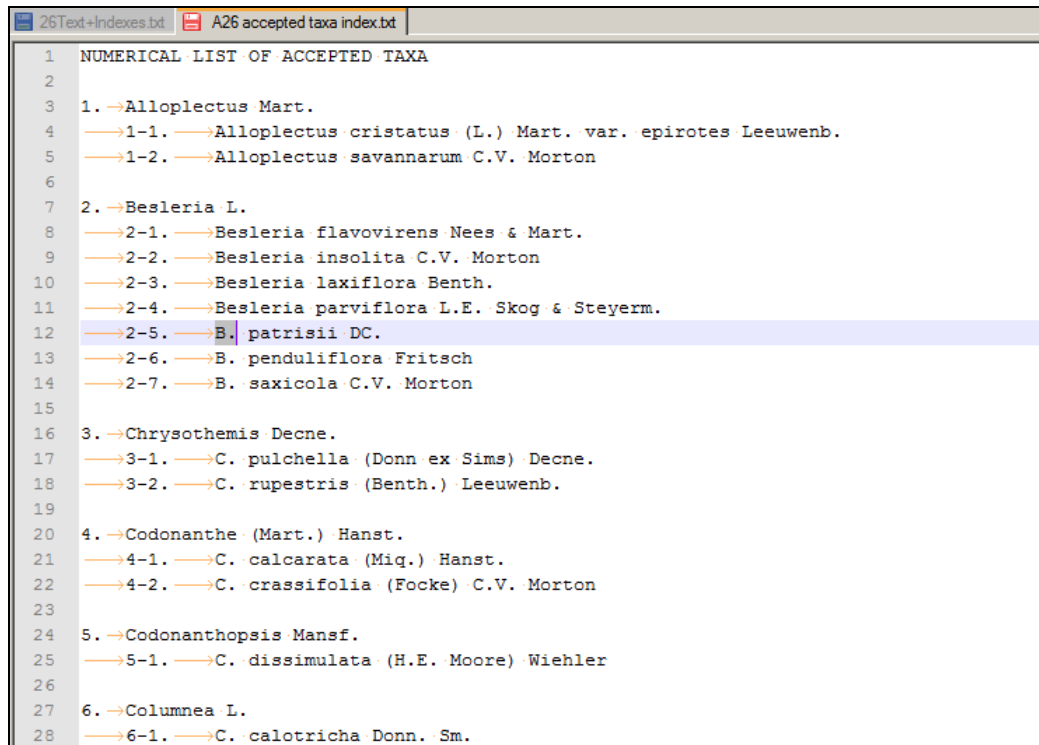


Figure 111: Expanding genus abbreviations into full genus names.

- 3) Now you have to go back to the Unicode-encoded (cleaned up) text file. Find the “Collections studied”-index.
 - a. Select the whole index *for a single family* and **copy** (not “cut”) it, using “Ctrl-C”.
 - b. Create another new file using “Ctrl-N”.
 - c. Paste the “Collections studied”-index into it, and save it with a meaningful name (Figure 112). In the example used there is only one family in the volume.
 - d. Now you will use Notepad++’ “Replace”-function to replace the numerical codes in the “Collections studied”-index by their corresponding taxon name from the “Numerical list of accepted taxa”-index.
 - i. First open up the “Replace” window by pressing “Ctrl-H”. Go to the file with the “Numerical list of accepted taxa”-index with expanded genera by clicking on its tab in the main Notepad++ window (Figure 113).

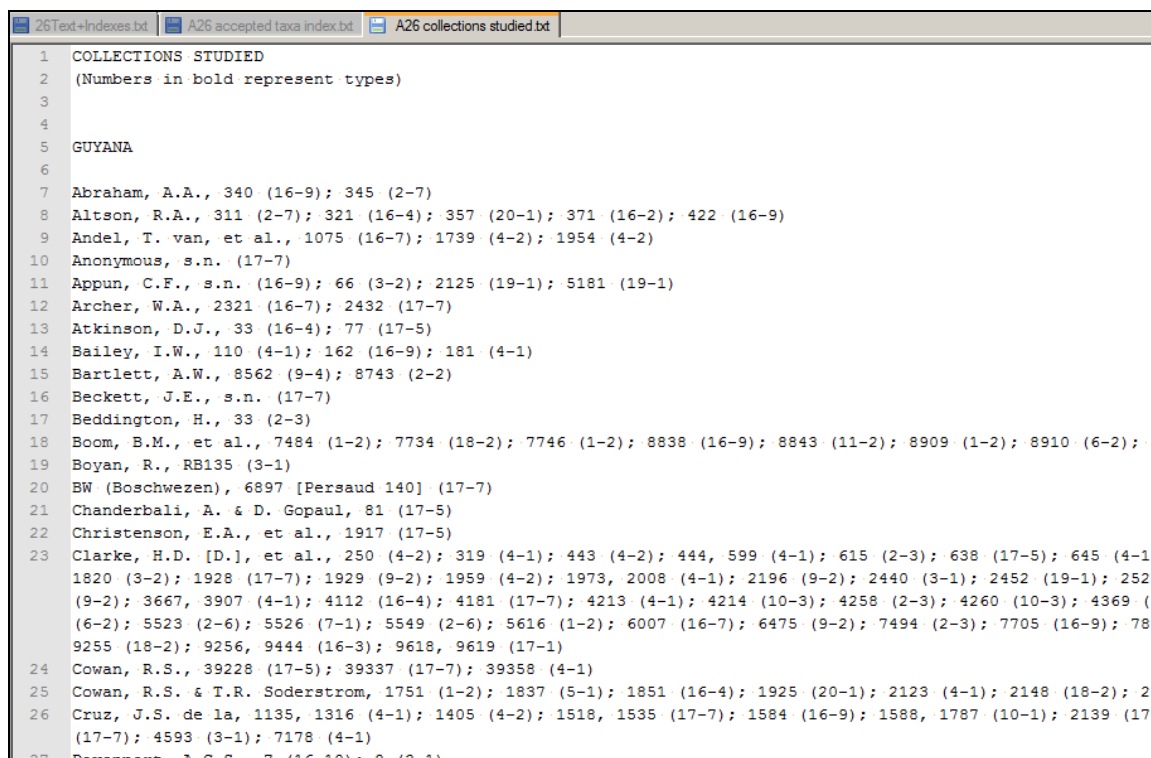


Figure 112: "Collections studied"-index in a new text file.

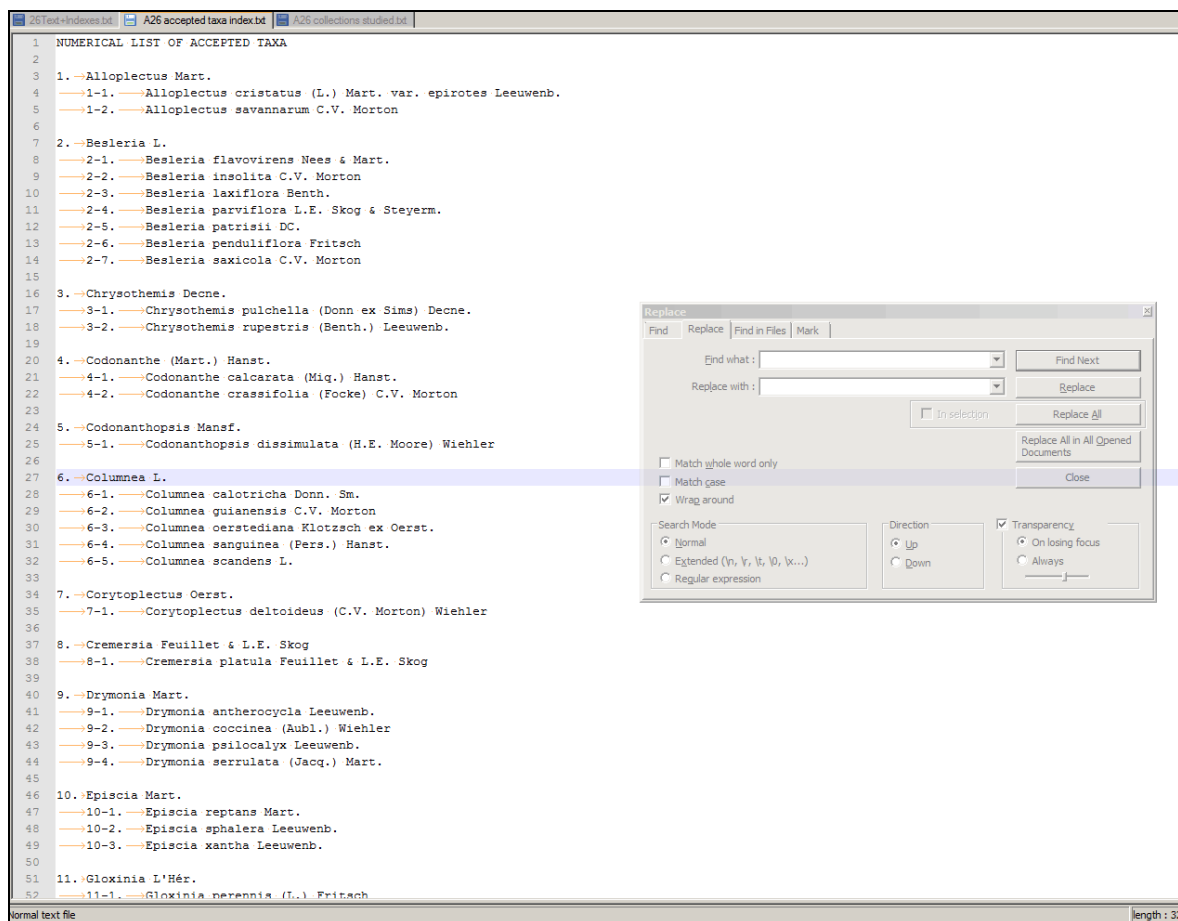


Figure 113: "Numerical list of accepted taxa"-index file with "Replace" window.

- ii. Because of practical reasons¹, it is easiest to start with the last taxon and work back in reverse order.
- iii. So go to the end of the file, select the numerical code for the last taxon, copy it, and paste it into the “Find what:” field of the “Replace”-window (Figure 114).

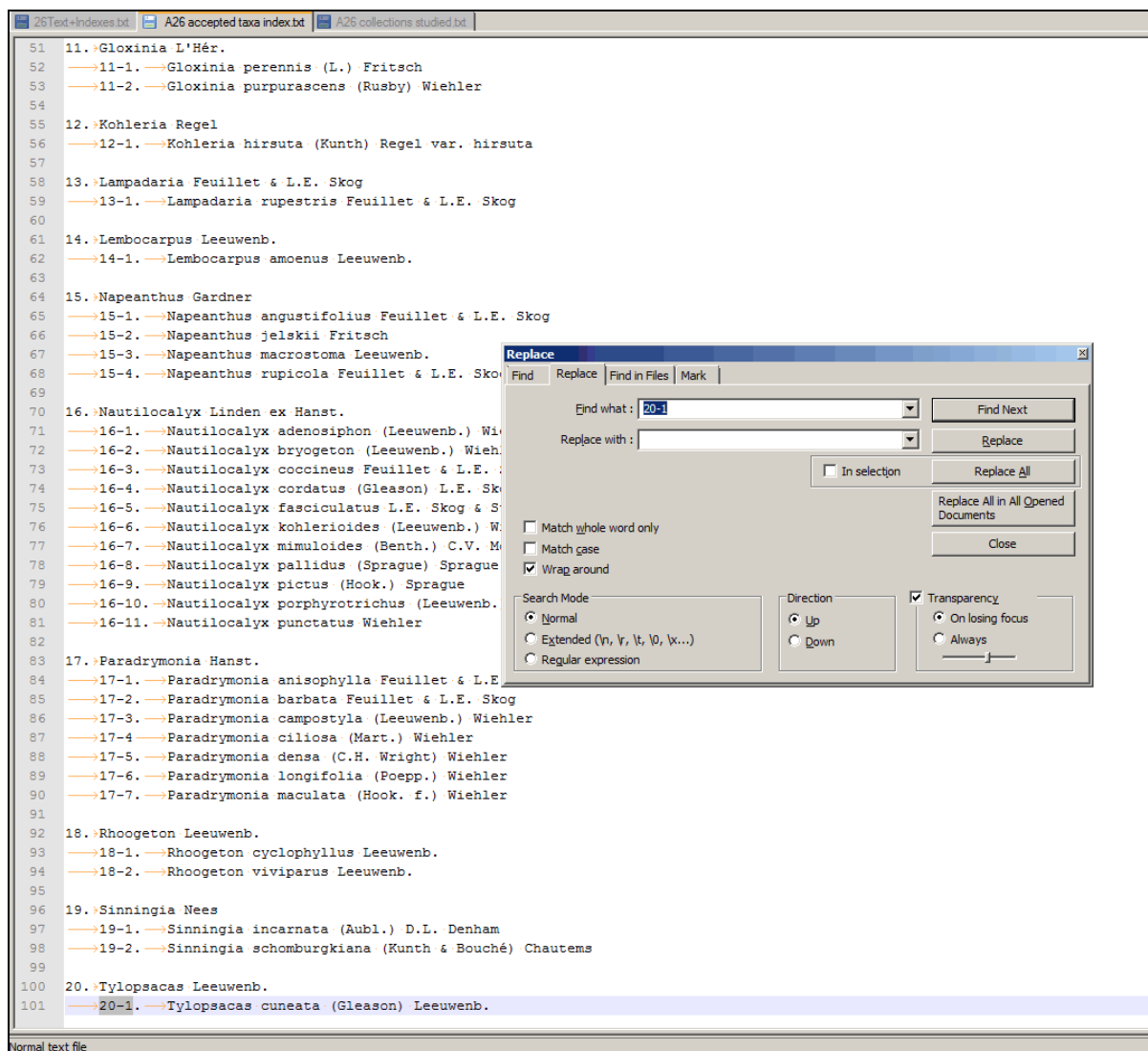


Figure 114: Copying the numerical code for the last taxon into the "Replace"-window.

- iv. Repeat this for the taxon name, but paste the taxon name into the “Replace with:”-field of the “Replace”-window. You can include the author names if you wish. Do not close the “Replace”-window.
- v. Now go to the file with the “Collections studied”-index by clicking on its tab in the main Notepad++ window. You should see something similar to Figure 115.

¹ More precisely, due to the method chosen a number like “1-1” has overlap with a number like “1-12”. This can be counteracted by the use of brackets in the “Replace”-window, but with a large number of taxa this quickly becomes rather cumbersome.

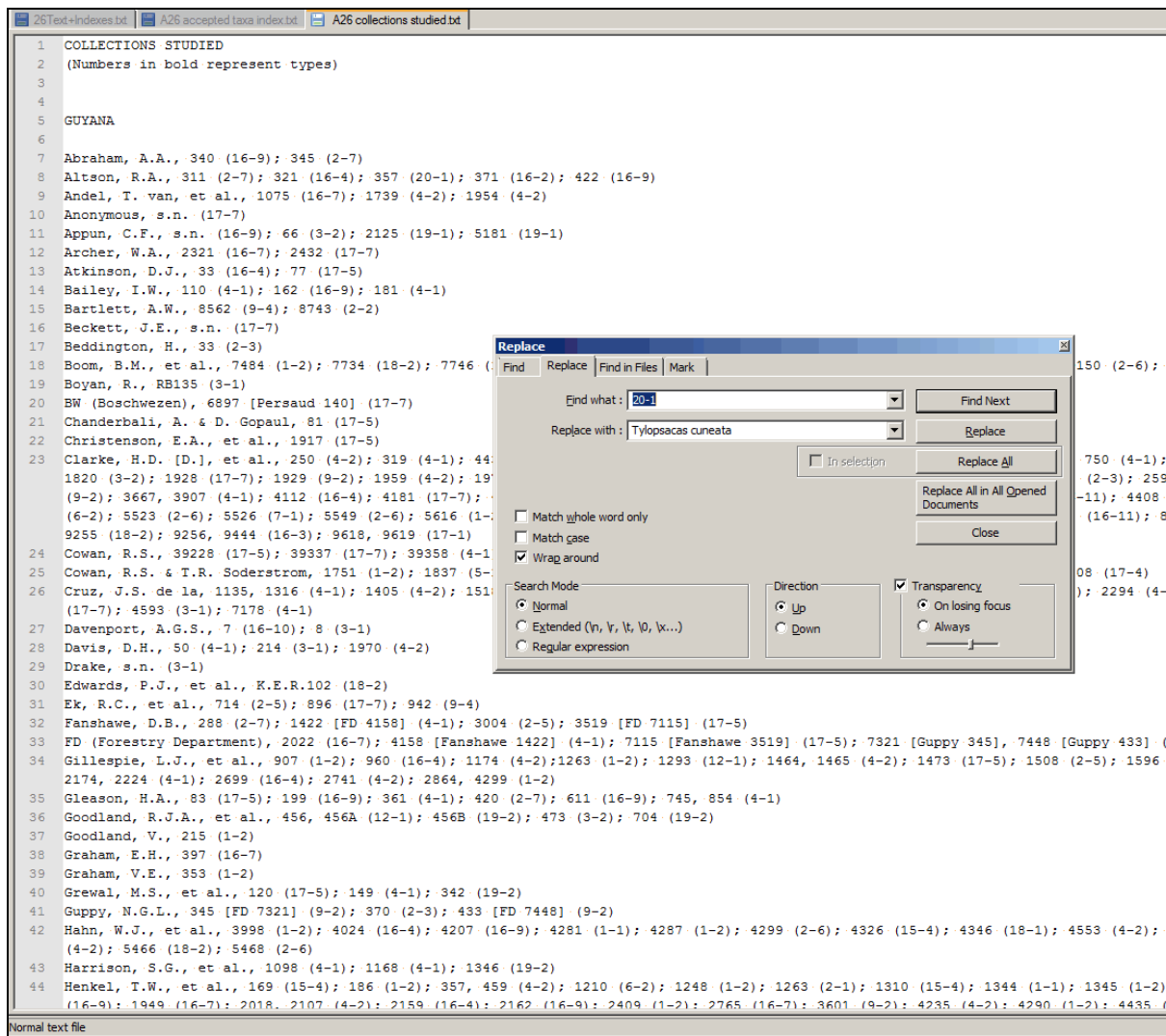


Figure 115: File with "Collections studied"-index and "Replace"-window.

- vi. Now you can click on the "Replace All"-button in the "Replace"-window to replace that particular numerical code with its corresponding taxon name everywhere in the file (Figure 116).
- vii. You will have to repeat steps 3d i-vi. for all of the other taxa of that family.
- e. Once done with a family, cut its "Collections studied"-index out of the file created in step 3b) and paste it over the original text for that family in the cleaned-up text (Figure 117).
- f. You will have to repeat steps 3a-e for each additional family in the volume. Take care to use the species names corresponding to each additional family every time.

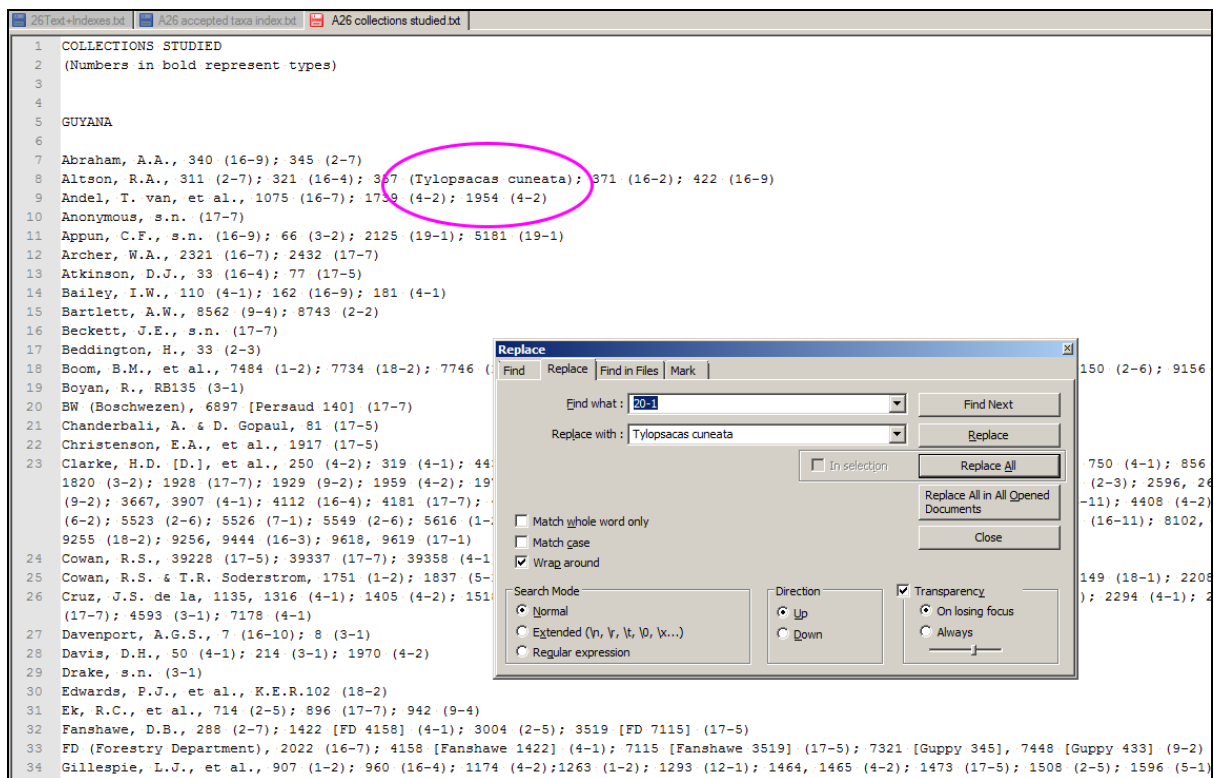


Figure 116: File with "Collections studied"-index with first numerical code replaced by corresponding taxon name.

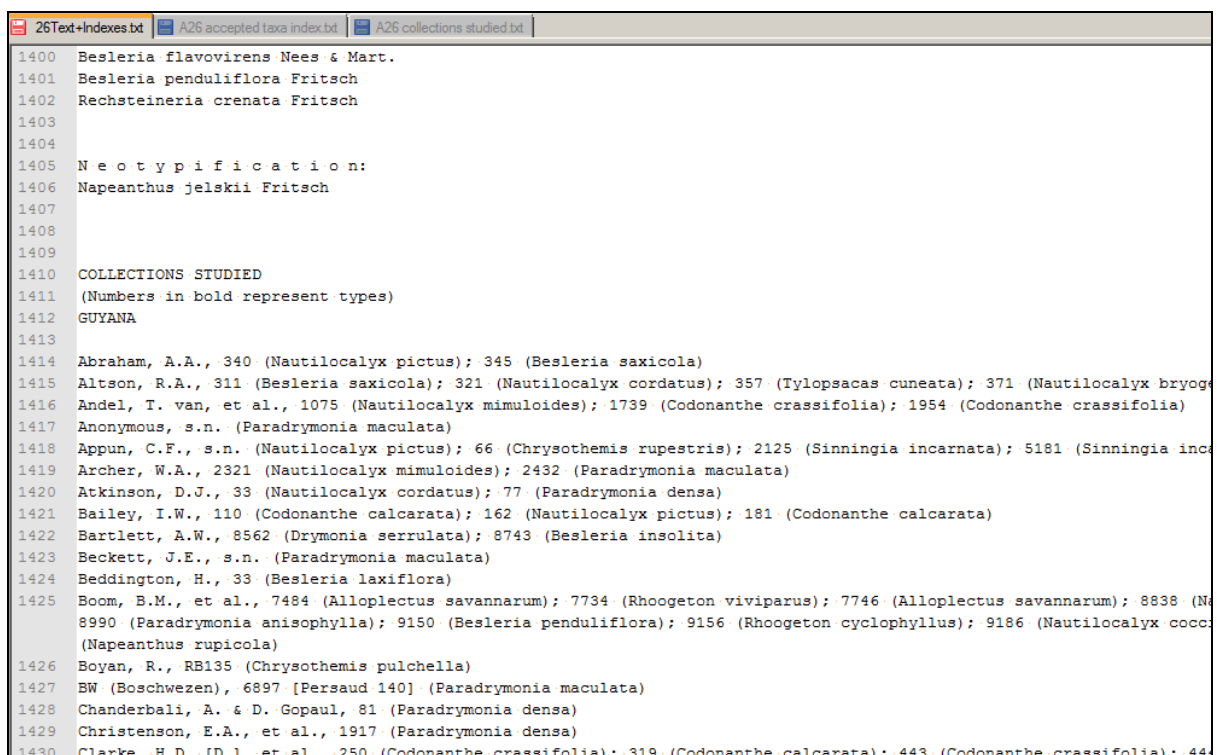


Figure 117: Cleaned up text file with "Collections studied"-index with taxon names inserted instead of numerical codes.

Moving “Collections studied”-index and wood descriptions

Finally, you have to move the “Collections studied”-index and eventually present wood descriptions from the end of the volume to the (corresponding) family description using cut and paste. Just paste them at the end of the correct family taxon treatment, before the key(s) to the lower taxa.

Now you can continue with the scripts for Flora of the Guianas, as described in Appendix I of **script use.doc**.