Preparing OCRd PDFs for use in Wordfast

or getting to know your source text from the inside out

By Abigail Clay
Freelance German > English pharma/medicine translator

Topics

- Intro
- Why convert PDFs?
- Convert a PDF job, yes or no?
- How to get editable text from dead PDFs
- Using Acrobat Pro to OCR dead PDFs
- Getting text from live PDFs
- Software and hardware overview
- Formatting Examples
- How to clear all formatting
- Workflow: Formatting
- How to remove unwanted graphics
- How to collect all graphics in a new file
- Find and replace
- Text Correction Examples
- Workflow: Text Correction
- How to remove tags
- How to record a macro
- How (not) to write a macro
- Formatting for alignment

Intro

- Translating professionally since 1991
- First translations done with a fax and Microsoft Word: formatting the document was a part of every job
- Next CAT tools with live text came along: formatting not necessary
- With OCR'd text: formatting and text prep are equally important

Why convert PDFs?

Freelancers

- Compare with client's stated word count
- You can use and add to your TMs and glossaries
- Control over text that is imported into Wordfast
- Overview of translation: you get the big picture, the context, which is harder to do with a TXML file alone
- Become more intimately acquainted with the text before you start translating; "pre-translation phase"
- Can preview the final document when and as you wish
- Can charge more for formatting
- You don't "give away" your TM to clients
- Can be used to align files and create editable glossaries

Why convert PDFs?

Project managers

- Ability to use (server-based) TMs/glossaries
- Ability to develop and manage client-specific TMs and glossaries
- Terminology extraction
- Control over segments/output etc. as for freelancer

Convert a PDF job, yes or no?

Which PDF jobs are worth OCRing?

- Material is something you translate a lot or for a regular client and usually only comes in the form of PDFs so it's worth making a personal template (e.g., ethics committee letters)
- Easy (enough) formatting (or you have a DTP department)
- Text is dark and light/dark contrast is good
- Little or no handwriting

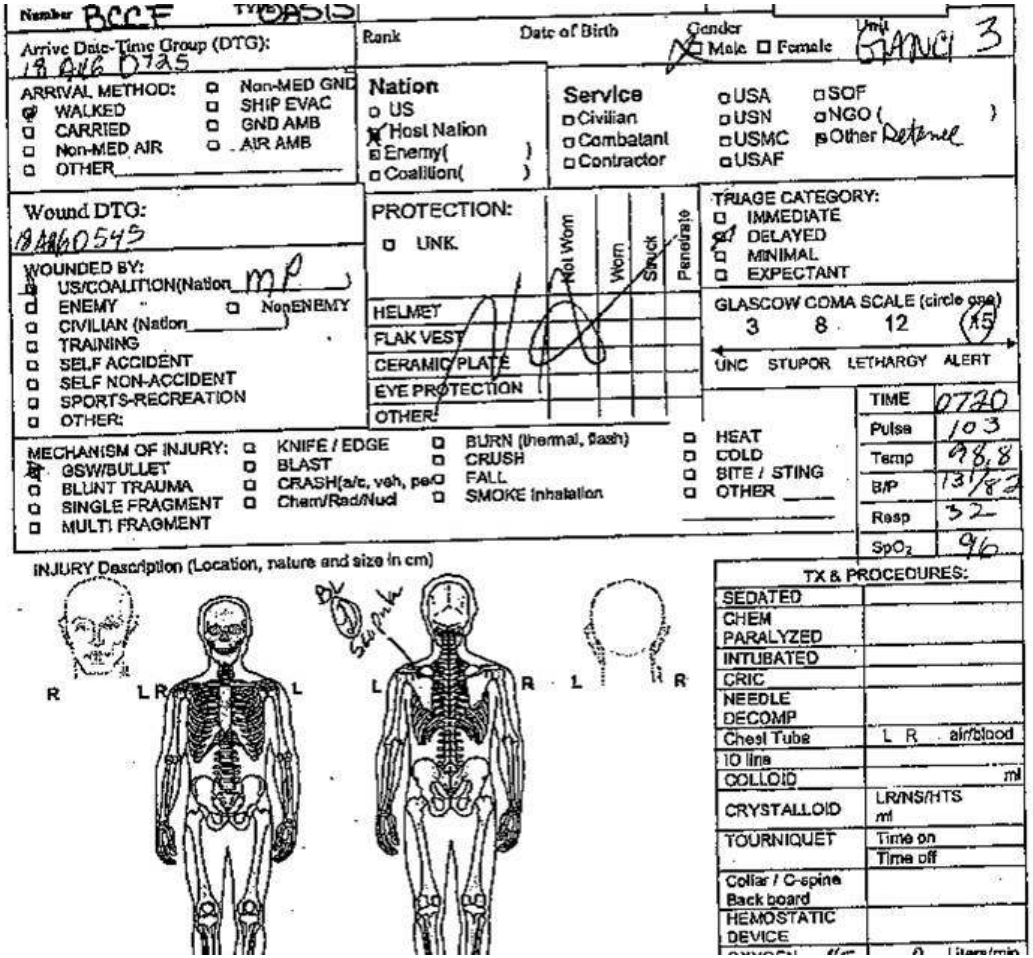
Convert a PDF job, yes or no?

Are you an OCR personality?

- Need to know Word (or whatever file type your client is requesting) layout and formatting well
- Need to understand how CAT tools perform segmentation
- Need Acrobat Reader Pro or other OCR software
- Helpful to like to experiment
- Helpful to view all formatting as making a template for future jobs (i.e., use tables)
- If you aren't comfortable asking to get paid for your prep time, then it often isn't worth it in terms of the hourly rate you get…
- …but may be in terms of the use of the TM/glossaries/formatting template in the future)

Would you OCR this document?

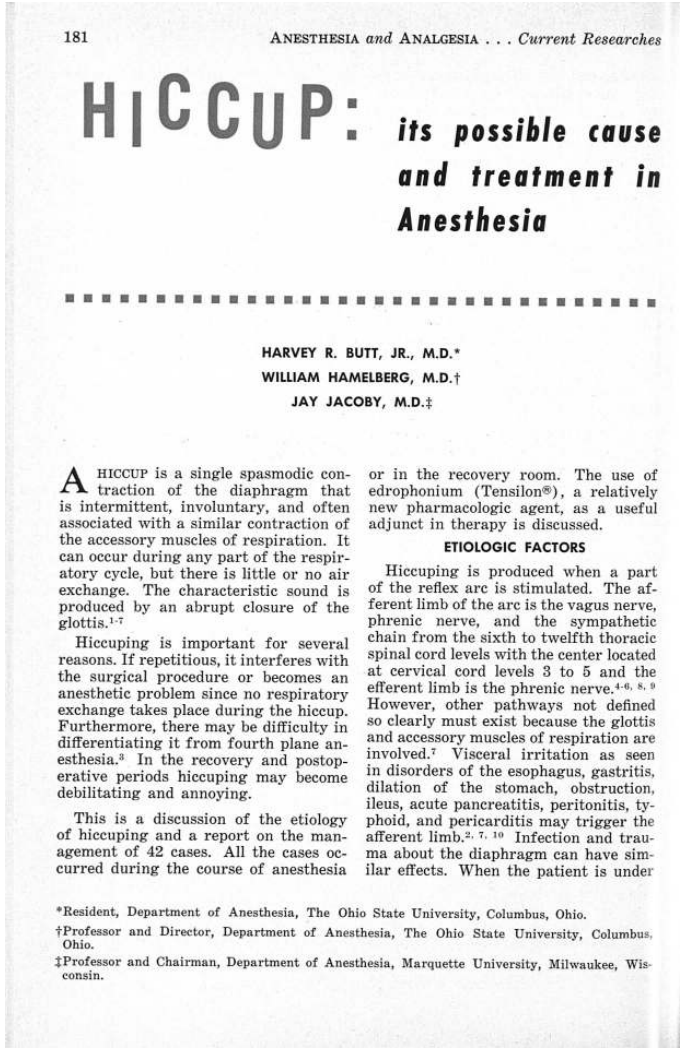
No, but I might make a template by hand, if I translated a lot of this type of report



Accept a PDF job, yes or no?

Would you OCR this document?

Yes, definitely.



181 ANESTHESIA *and* ANALGESIA . . . *Current Researches*

HICCUP: *its possible cause and treatment in Anesthesia*

HARVEY R. BUTT, JR., M.D.*
WILLIAM HAMELBERG, M.D.†
JAY JACOBY, M.D.‡

A HICCUP is a single spasmodic contraction of the diaphragm that is intermittent, involuntary, and often associated with a similar contraction of the accessory muscles of respiration. It can occur during any part of the respiratory cycle, but there is little or no air exchange. The characteristic sound is produced by an abrupt closure of the glottis.[1-7]

Hiccuping is important for several reasons. If repetitious, it interferes with the surgical procedure or becomes an anesthetic problem since no respiratory exchange takes place during the hiccup. Furthermore, there may be difficulty in differentiating it from fourth plane anesthesia.[8] In the recovery and postoperative periods hiccuping may become debilitating and annoying.

This is a discussion of the etiology of hiccuping and a report on the management of 42 cases. All the cases occurred during the course of anesthesia or in the recovery room. The use of edrophonium (Tensilon®), a relatively new pharmacologic agent, as a useful adjunct in therapy is discussed.

ETIOLOGIC FACTORS

Hiccuping is produced when a part of the reflex arc is stimulated. The afferent limb of the arc is the vagus nerve, phrenic nerve, and the sympathetic chain from the sixth to twelfth thoracic spinal cord levels with the center located at cervical cord levels 3 to 5 and the efferent limb is the phrenic nerve.[4-6, 8, 9] However, other pathways not defined so clearly must exist because the glottis and accessory muscles of respiration are involved.[7] Visceral irritation as seen in disorders of the esophagus, gastritis, dilation of the stomach, obstruction, ileus, acute pancreatitis, peritonitis, typhoid, and pericarditis may trigger the afferent limb.[2, 7, 10] Infection and trauma about the diaphragm can have similar effects. When the patient is under

*Resident, Department of Anesthesia, The Ohio State University, Columbus, Ohio.
†Professor and Director, Department of Anesthesia, The Ohio State University, Columbus, Ohio.
‡Professor and Chairman, Department of Anesthesia, Marquette University, Milwaukee, Wisconsin.

How to get editable text from *dead* PDFs

- Acrobat Pro

  Very easy workflow, but very limited settings
- OCR using software such as ABBYY Finereader or OmniPage Pro

  More time-consuming and complicated workflow, but can select which areas of a document you want to OCR and you can train the OCR engine to perform better the more texts you do
- Google Drive (for up to 3 pages)

Using Acrobat Pro to OCR *dead* PDFs

- Performing OCR on dead PDFs with Acrobat Pro is as easy as just saving the file as an editable format such as docx, rtf, txt
- No one output format is the best for all documents—you many have to experiment to see if RTF, DOC or TXT gives you the best result

Using Acrobat Pro to OCR *dead* PDFs

- You can customize OCR settings to indicate the source language and whether or not to include images
- You can also have Acrobat just perform OCR (but not output as a different file format) so you can have a searchable file (Tools > Optimize Scanned PDF)

Getting text from *live* PDFs

- There is a lot of software out there that extracts text from live PDFs
- Free ones tend to have limits on number of pages or output formats
- Prefer to either use Acrobat or copy/paste and use TransTools for paragraph returns mid-sentence

Getting text from *live* PDFs

PDF letter with live text

Which method is best?

1. Save as docx using Acrobat
2. Save as RTF using Acrobat
3. Copy and paste into Word
4. Use directly with CAT tool
5. Print to jpeg, then print to PDF, save as rtf
6. Print a hard copy, scan and OCR with Acrobat

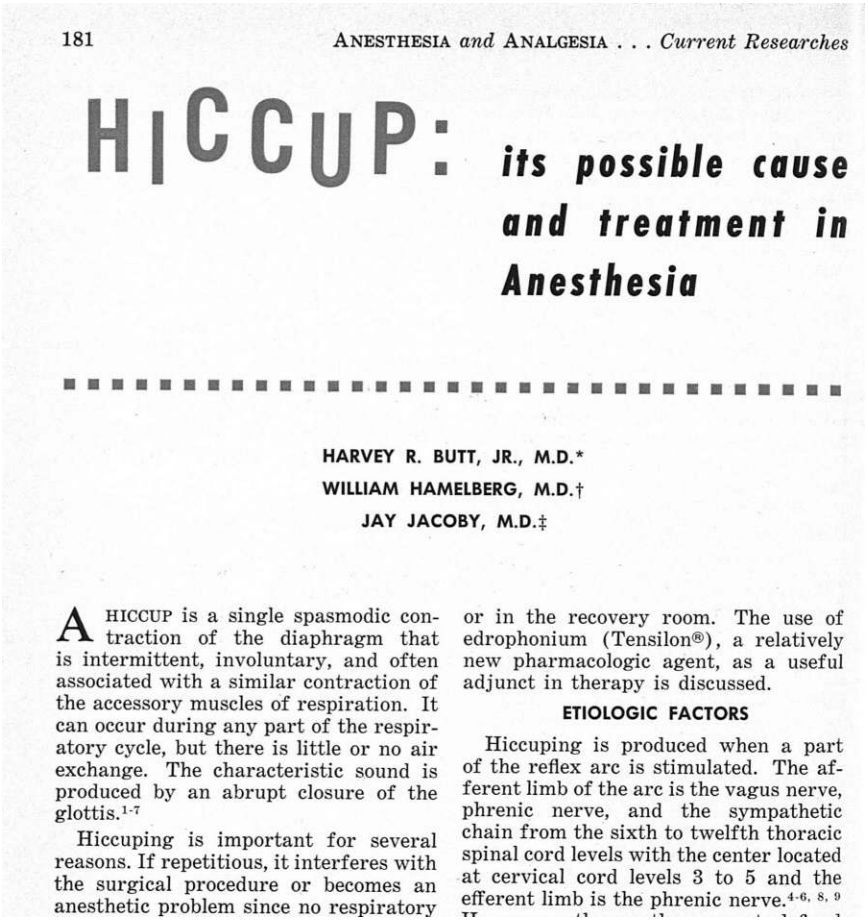
3, with 5 as a close second

Software overview

- OCR
  - Adobe Acrobat Professional (ca. €100)
  - ABBYY Finereader Pro 11 (ca. € 129)
  - OmniPage Pro (ca. €120- € 199)
- TransTools (Word add-in) (full version €25)
- CodeZapper (Word add-in) (€20)

Formatting Examples

Dead PDF original



181 ANESTHESIA *and* ANALGESIA . . . *Current Researches*

HICCUP: its possible cause and treatment in Anesthesia

HARVEY R. BUTT, JR., M.D.*
WILLIAM HAMELBERG, M.D.†
JAY JACOBY, M.D.‡

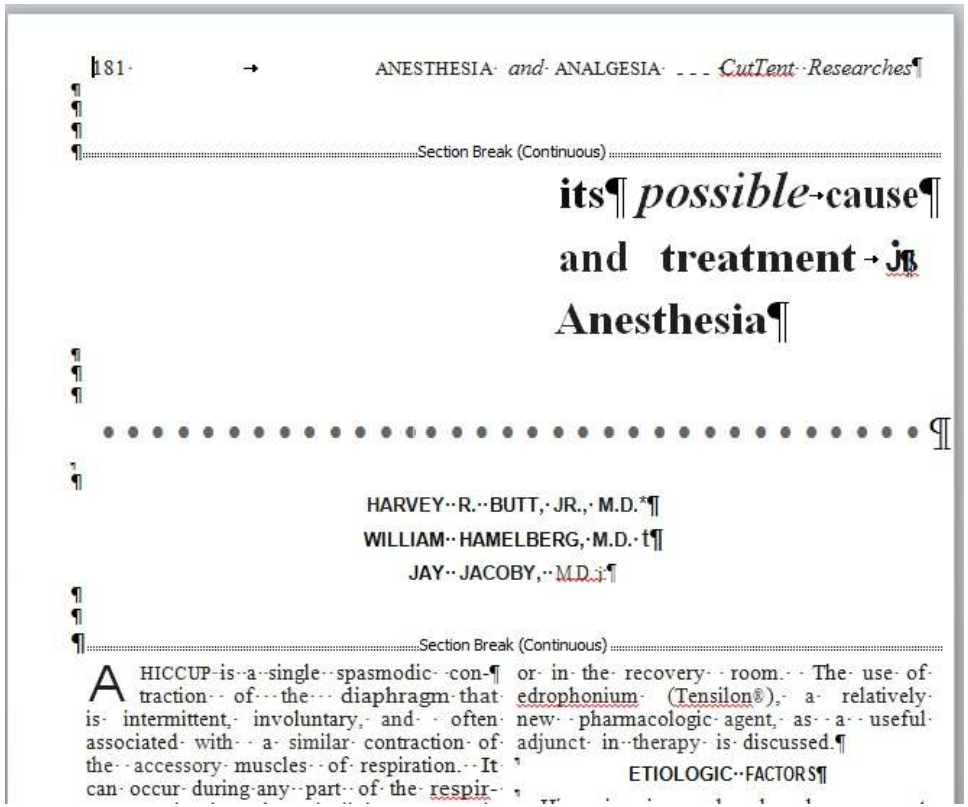
A HICCUP is a single spasmodic contraction of the diaphragm that is intermittent, involuntary, and often associated with a similar contraction of the accessory muscles of respiration. It can occur during any part of the respiratory cycle, but there is little or no air exchange. The characteristic sound is produced by an abrupt closure of the glottis.[1-7]

Hiccuping is important for several reasons. If repetitious, it interferes with the surgical procedure or becomes an anesthetic problem since no respiratory or in the recovery room. The use of edrophonium (Tensilon®), a relatively new pharmacologic agent, as a useful adjunct in therapy is discussed.

ETIOLOGIC FACTORS

Hiccuping is produced when a part of the reflex arc is stimulated. The afferent limb of the arc is the vagus nerve, phrenic nerve, and the sympathetic chain from the sixth to twelfth thoracic spinal cord levels with the center located at cervical cord levels 3 to 5 and the efferent limb is the phrenic nerve.[4-6, 8, 9]

Saved as DOCX from Acrobat



181 → ANESTHESIA and ANALGESIA . . . CutTent Researches¶

Section Break (Continuous)

its¶ *possible* cause¶
and treatment in¶
Anesthesia¶

HARVEY R. BUTT, JR., M.D.*¶
WILLIAM HAMELBERG, M.D. †¶
JAY JACOBY, M.D. ‡¶

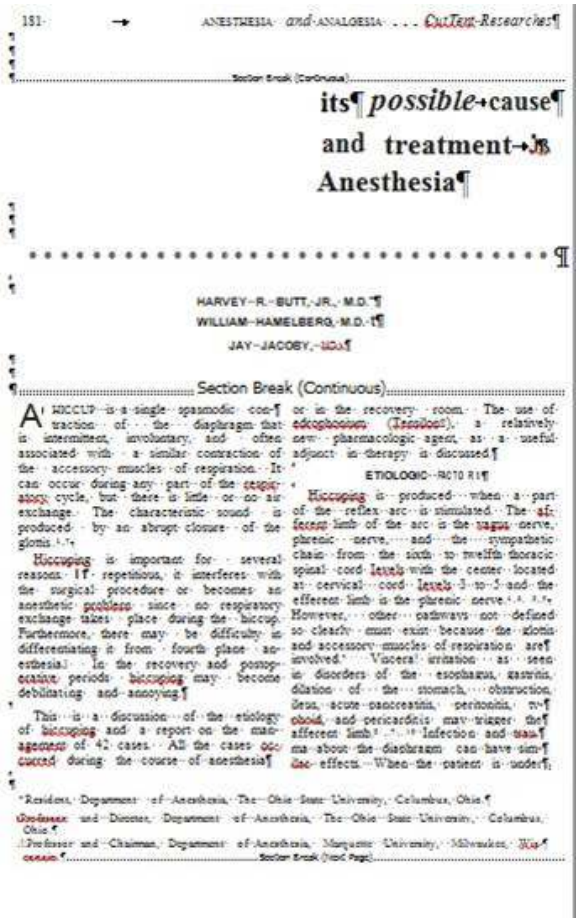
Section Break (Continuous)

A HICCUP is a single spasmodic con-¶ traction of the diaphragm that is intermittent, involuntary, and often associated with a similar contraction of the accessory muscles of respiration. It can occur during any part of the respir- or in the recovery room. The use of edrophonium (Tensilon®), a relatively new pharmacologic agent, as a useful adjunct in therapy is discussed.¶

ETIOLOGIC FACTORS¶

Formatting Examples

Unformatted



181 → ANESTHESIA and ANALGESIA . . . CutTent-Researches

its possible cause
and treatment
Anesthesia

HARVEY R. BUTT, JR., M.D.
WILLIAM HAMELBERG, M.D.
JAY JACOBY, M.D.

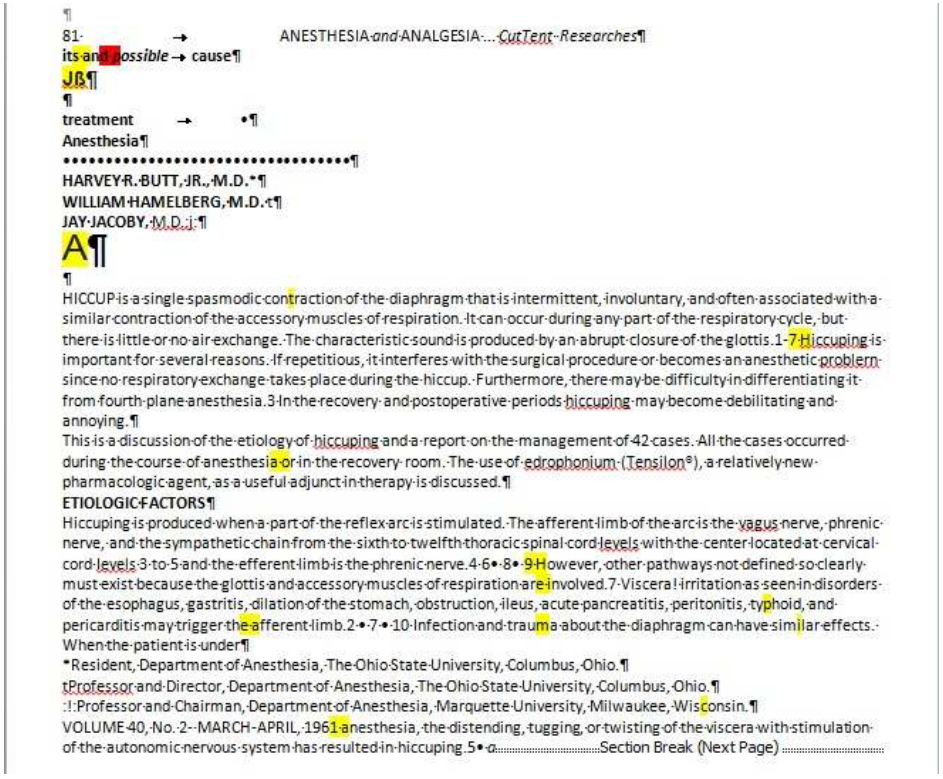
Section Break (Continuous)

A HICCUP is a single spasmodic contraction of the diaphragm that is intermittent, involuntary, and often associated with a similar contraction of the accessory muscles of respiration. It can occur during any part of the respiratory cycle, but there is little or no air exchange. The characteristic sound is produced by an abrupt closure of the glottis.

Hiccuping is important for several reasons. If repetitious, it interferes with the surgical procedure or becomes an anesthetic problem since no respiratory exchange takes place during the hiccup. Furthermore, there may be difficulty in differentiating it from fourth plane anesthesia. In the recovery and postoperative periods hiccuping may become debilitating and annoying.

This is a discussion of the etiology of hiccuping and a report on the management of 42 cases. All the cases occurred during the course of anesthesia

After running macros, TransTools, F/R



81 → ANESTHESIA and ANALGESIA . . . CutTent Researches
its and possible → cause
JB
treatment → •
Anesthesia
HARVEY R. BUTT, JR., M.D.*
WILLIAM HAMELBERG, M.D.†
JAY JACOBY, M.D., ‡
A

HICCUP is a single spasmodic contraction of the diaphragm that is intermittent, involuntary, and often associated with a similar contraction of the accessory muscles of respiration. It can occur during any part of the respiratory cycle, but there is little or no air exchange. The characteristic sound is produced by an abrupt closure of the glottis. 1-7 Hiccuping is important for several reasons. If repetitious, it interferes with the surgical procedure or becomes an anesthetic problem since no respiratory exchange takes place during the hiccup. Furthermore, there may be difficulty in differentiating it from fourth plane anesthesia. 3 In the recovery and postoperative periods hiccuping may become debilitating and annoying.

This is a discussion of the etiology of hiccuping and a report on the management of 42 cases. All the cases occurred during the course of anesthesia or in the recovery room. The use of edrophonium (Tensilon®), a relatively new pharmacologic agent, as a useful adjunct in therapy is discussed.

ETIOLOGIC FACTORS

Hiccuping is produced when a part of the reflex arc is stimulated. The afferent limb of the arc is the vagus nerve, phrenic nerve, and the sympathetic chain from the sixth to twelfth thoracic spinal cord levels with the center located at cervical cord levels 3 to 5 and the efferent limb is the phrenic nerve. 4-6 • 8 • 9 However, other pathways not defined so clearly must exist because the glottis and accessory muscles of respiration are involved. 7 Visceral irritation as seen in disorders of the esophagus, gastritis, dilation of the stomach, obstruction, ileus, acute pancreatitis, peritonitis, typhoid, and pericarditis may trigger the afferent limb. 2 • 7 • 10 Infection and trauma about the diaphragm can have similar effects. When the patient is under

*Resident, Department of Anesthesia, The Ohio State University, Columbus, Ohio.
†Professor and Director, Department of Anesthesia, The Ohio State University, Columbus, Ohio.
‡Professor and Chairman, Department of Anesthesia, Marquette University, Milwaukee, Wisconsin.

VOLUME 40, No. 2 MARCH-APRIL, 1961 anesthesia, the distending, tugging, or twisting of the viscera with stimulation of the autonomic nervous system has resulted in hiccuping. 5 • 6 Section Break (Next Page)

How to clear all formatting

- Notepad on Windows computers
- Paste into Word using the “Keep Text Only” option
- Use “Clear formatting” button in Word

How to clear all formatting

181· → ANESTHESIA·and·ANALGESIA·... CutTent··Researches¶
·¶
its¶
and¶
·¶
possible → cause¶
treatment → •¶
·¶
Anesthesia¶
¶
••••••••••••••••••••••••••••••••••¶
HARVEY··R.··BUTT,·JR.,·M.D.*¶
WILLIAM··HAMELBERG,·M.D.†¶
JAY··JACOBY,··M.D.‡¶
·¶
HICCUP··is··a··single··spasmodic··con-¶
traction··of···the···diaphragm·that·is·intermittent,· involuntary,·and··often·associated·with··a·similar·contraction·of·the··accessory·muscles·of·respiration.··It·can·occur·during·any·part·of·the·respir··atory·cycle,·but··there·is·little··or··no·air·exchange.·The·characteristic·sound··is·produced··by·an·abrupt·closure··of·the·glottis.1-7¶
Hiccuping· is·important·for··several·reasons.·If·repetitious,·it·interferes·with·the·surgical·procedure·or·becomes·an·anesthetic·problem·since··no·respiratory·exchange·takes··place·during·the···hiccup.·Furthermore,·there·may··be·difficulty·in·differentiating··it·from··fourth·plane··an··esthesia.3···In·the·recovery·and·postop··erative·periods··hiccuping·· may··become·debilitating··and··annoying.¶
¶
This···is··a··discussion·· ··of···the···etiology·of·hiccuping·· and·a·report·on·the·man··agement·of·42·cases.··All·the·cases·oc··curred·during·the···course···of···anesthesia¶
·¶
or·in·the·recovery··room.··The·use·of·edrophonium·· (Tensilon®),··a·relatively·new··pharmacologic·· agent,·as··a··useful·adjunct·· in···therapy·is·discussed.¶
¶
ETIOLOGIC·FACTORS¶
¶
Hiccuping·· is··produced·· ··when··a···part·of·the··reflex···arc···is·stimulated.·· ·The···af··ferent·limb··of·the·arc·is·the·vagus··nerve,·phrenic·· ··nerve,·····and·····the·····sympathetic·chain··from···the·sixth···to·twelfth·thoracic·spinal··cord·levels·with·the·center···located·at···cervical···cord···levels··3···to··5···and···the·efferent·limb·is·the···phrenic·nerve.4·6••8••9¶
However,···other···pathways···not···defined·so·clearly···must···exist···because···the···glottis·and·accessory···muscles·of·respiration·are¶
involved.7·····Viscera!·irritation···as···seen·in·disorders·of·the··esophagus,·gastritis,·dilation··of···the···stomach,·····obstruction,·· ileus,···acute··pancreatitis,···peritonitis,·· ty-¶
phoid,·· ·and···pericarditis·may···trigger·the¶
afferent·limb.2·•···7·•···10·····Infection·· and··trau-¶
ma···about···the···diaphragm·can···have·sim-¶
ilar···effects.···When···the···patient·is···under¶
¶
*Resident,···Department···of···Anesthesia,···The···Ohio···State···University,···Columbus,···Ohio.¶
†Professor···and···Director···Department···of···Anesthesia···The···Ohio···State···University···Columbus···Ohio.¶

Workflow: Formatting

1. Set paper size, margins, single column, set proofing language as the source language (recorded macro)
2. Font set to Calibri 10, normal, normal, 100%, paragraph no space before or after, single spacing (recorded macro)

Workflow: Formatting

1. Remove all lines (TransTools)
2. Take all text out of text boxes (TransTools)
3. Clean up tables (TransTools)
4. Remove incorrect paragraph breaks (¶), especially from text converted from live/native PDFs. (TransTools, manually)
5. Make all graphics inline (doesn't always get them all) (TransTools)

Workflow: Formatting

1. Remove all text boxes (macro from VBA code)
2. Remove all graphics and shapes (macro from VBA code)
3. Put header/footer text in a header or footer so you don't have to recheck this text for every page—OCR will not do that for you.

How to remove unwanted graphics

Manual

- Select All to highlight entire document (graphics will have a dark frame around them)
- Manually remove each graphic by clicking on it and deleting

Find and replace (for all inline graphics)

- Open Find/Replace dialogue box and in the Find box click Special > Graphic (^g)
- In the Replace box put nothing; replace all

How to remove unwanted graphics

- Macro for floating graphics

```
Sub demo()
    Dim oShp As Shape
    Dim oIShp As InlineShape
    For Each oShp In
ActiveDocument.Shapes
        oShp.Delete
    Next
    For Each oIShp In
ActiveDocument.InlineShap
es
        oIShp.Delete
    Next
End Sub
```

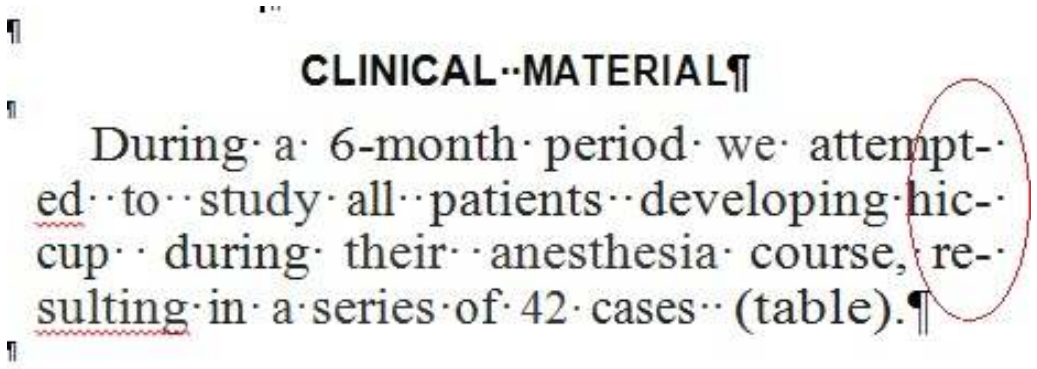
How to collect all graphics in a new file

- TransTools

  Has a nifty feature of being able to pull all (inline) graphics from a document into new file to use later (TransTools tab, TransTools button, Special, Collect All Graphic From Active Document)

Find and Replace

- Ctrl + H shortcut opens dialogue box
- Don’t forget you can replace formatting as well character strings
- Soft hyphens: copy and paste into Find box (replace with nothing)



soft hyphens

Text Correction Example

its *possible* cause and treatment in Anesthesia

HARVEY R. BUTT, JR., M.D.*
WILLIAM HAMELBERG, M.D.†
JAY JACOBY, M.D.‡

Section Break (Continuous)

A HICCUP is a single spasmodic contraction of the diaphragm that is intermittent, involuntary, and often associated with a similar contraction of the accessory muscles of respiration. It can occur during any part of the respiratory cycle, but there is little or no air exchange. The characteristic sound is produced by an abrupt closure of the glottis.[1-7]

Hiccuping is important for several reasons. If repetitious, it interferes with the surgical procedure or becomes an anesthetic problem since no respiratory exchange takes place during the hiccup. Furthermore, there may be difficulty in differentiating it from fourth plane anesthesia.[3] In the recovery and postoperative periods hiccuping may become debilitating and annoying.

This is a discussion of the etiology of hiccuping and a report on the management of 42 cases. All the cases occurred during the course of anesthesia or in the recovery room. The use of edrophonium (Tensilon®), a relatively new pharmacologic agent, as a useful adjunct in therapy is discussed.

ETIOLOGIC FACTORS

Hiccuping is produced when a part of the reflex arc is stimulated. The afferent limb of the arc is the vagus nerve, phrenic nerve, and the sympathetic chain from the sixth to twelfth thoracic spinal cord levels with the center located at cervical cord levels 3 to 5 and the efferent limb is the phrenic nerve.[1,2,4,7,8] However, other pathways not defined so clearly must exist because the glottis and accessory muscles of respiration are involved.[7] Visceral irritation as seen in disorders of the esophagus, gastritis, dilation of the stomach, obstruction, ileus, acute pancreatitis, peritonitis, typhoid, and pericarditis may trigger the afferent limb.[2,7,10] Infection and trauma about the diaphragm can have similar effects. When the patient is under anesthesia, the distending, tugging, or twisting of the viscera with stimulation of the autonomic nervous system has resulted in hiccuping.[3,6]

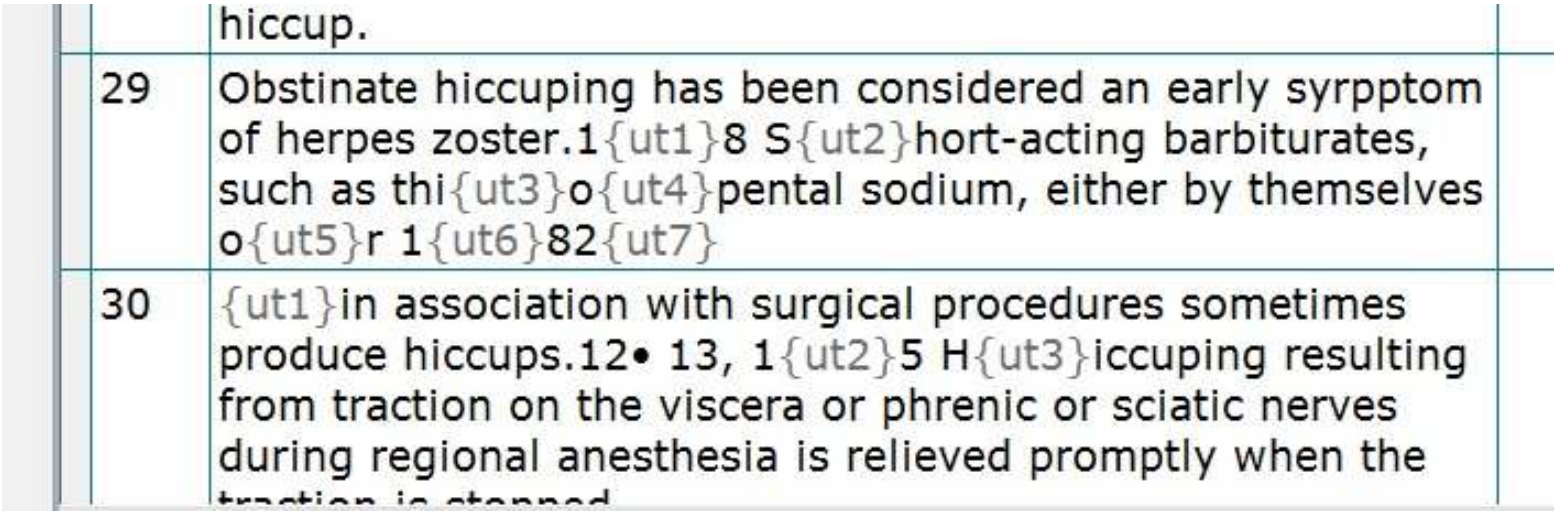
*Resident, Department of Anesthesia, The Ohio State University, Columbus, Ohio.
†Professor and Director, Department of Anesthesia, The Ohio State University, Columbus, Ohio.
‡Professor and Chairman, Department of Anesthesia, Marquette University, Milwaukee, Wisconsin.

Workflow: Text Correction

1. Go through document from the beginning and correct text manually (you can do this in ABBYY FineReader or OmniPage Pro before exporting to a Word document, but initially it is very time-consuming)
2. Use F/R for common OCR errors (e.g., hn instead of m)
3. Look at numbers and symbols, check for superscript and subscript (Ctrl+Shift+ +; Ctrl+Shift+=)
4. Change font to a serif typeface like Times New Roman so you can see if 1 and lowercase L were OCRd correctly
5. Check for *soft* hyphens and delete with F/R if necessary by copying hyphen and pasting in Find box
6. You will probably have to retype some sections of dead PDFs.

How to remove tags

- Clear all formatting first using Word or Notepad
- Use TransTools
- Use CodeZapper



	hiccup.
29	Obstinate hiccuping has been considered an early syrpptom of herpes zoster.1{ut1}8 S{ut2}hort-acting barbiturates, such as thi{ut3}o{ut4}pental sodium, either by themselves o{ut5}r 1{ut6}82{ut7}
30	{ut1}in association with surgical procedures sometimes produce hiccups.12• 13, 1{ut2}5 H{ut3}iccuping resulting from traction on the viscera or phrenic or sciatic nerves during regional anesthesia is relieved promptly when the

How to record a macro

- A macro is a series of commands and instructions that you group together as a single command to accomplish a task automatically.
- You need to have the Developer tab shown on your ribbon to record macros (File > Options > Customize ribbon)
- You cannot use a mouse to select text when recording a macro. You must use the keyboard to select text (e.g., Ctrl + A to select all text on Windows PCs)

How to record a macro

- When you name macros use underscore between words, e.g., Delete_all_graphics

Tip: Don't record one macro with all the actions you want to do—break the actions up into groups and run separately

How (not) to write a macro

Want to try your hand at using VBA code for tasks you can record in a macro?

- Hit Alt+F11 or Tools, Developer, Visual Basic.
- Learn Visual Basic for Applications ☺ *or*
- Copy useful code written by developers who have had the same problem you have had (e.g., http://www.vbforums.com/showthread.php?459263-deleting-all-graphics-in-Word-using-a-macro)
- Basic VBA macro tutorial: https://msdn.microsoft.com/en-us/library/office/ff604039%28v=office.14%29.aspx

How (not) to write a macro

- Open the Visual Basic editor (Alt + F11), Insert Module
- Paste code
- F5 to run macro to test it
- Can assign macro to button or shortcut keys
- Save to a new project or to Normal

Formatting for alignment

- Formatting text and layout is different for alignment
- For most alignment software, the key is that there be the same number of segments in the source and target languages
- Formatting is much less important: you can have a segment followed by a hard return and that's it.

Key Points

- Be careful about which documents you choose to OCR
- Keep your eye on the prize: text you can leverage in Wordfast, formatting that can be reused
- Control over your process and the translation from start to finish
- Invest in Acrobat Pro and TransTools
- No single technique will get you perfectly formatted and error-free text
- Keep macros limited to a group of tasks—not all source documents need the same formatting, so even though you will have to run more than one macro, you control the process better

“Love words, agonize over sentences. And pay attention to the world.”

– Susan Sontag

Contact

Abigail Clay, DE>EN pharma/medicine translator
www.germanandlanguagearts.com
abby@germanandlanguagearts.com
Skype: abbyclay