University of Zurich[UZH]

Institute of Computational Linguistics

Writing a Scientific Thesis

Mathias Müller

Contents

Words —– possession of the beggar and the king, obsession of the genius and the fool.
Anders Edenroth, *The Real Group*

1 Introduction

Writing your thesis should be an enjoyable experience. And yet, many people find rather little joy in it. It might be because wrestling with any sizeable project will put you through serious hardship at some point. But also, it might be because nobody told them what and how to write. Your courses on higher-order logic will not help here.

This is exactly the gap that this guide is trying to fill. Writing scientifically is not just a matter of finding the right words that will, hopefully, persuade your audience. An audience of scientists can only be persuaded with facts that result from a well-thought-out experiment. There are three crucial aspects to an experiment that is well thought-out. Firstly, you must motivate it by stating the problem you are trying to solve. Secondly, you must ensure that anyone can *reproduce* your results with the information you gave them. Thirdly, you must give an *interpretation* of the results, explaining what they mean. Chapter 2 is about the content of your thesis.

In order to live up to those expectations, you definitely need to plan your writing process. For once, there must be a method to the madness that drives creativity. Most importantly, you should think about the *order* in which you assemble bits and pieces of your thesis. Therefore, Chapter 3 is about *when* to do stuff in what *order*.

When it is clear to you what the content of your thesis will be and how you should organise yourself, it's time to think about the *form* your writing will take. If you read the second paragraph above very closely, you will notice that "writing scientifically is *not just* a matter of finding the right words". Which, of course, means that it is *also* a matter of finding the right words, it's just that stylistic considerations come last in the writing process. Last but not least! An eloquent style of writing is still important to get your meaning across. Chapter 4 will instruct you to write clearly and concisely.

Finally, Chapter 5 has more general advice for you. Chapter 6 lists things that this guide does not cover (yet), but that you probably also have to think about.

Much of this guide is stolen shamelessly from Peat (2002), Katz (2009) and the valuable teaching materials of Simon Clematide. If you find something useful in it, it is most likely taken from them and all the errors in it are mine.

2 Content

This section is about *what* to write about in your thesis. Describing the contents of your work is by far the easiest and straightforward topic of this guide. Scientific theses contain the following: 1) a knowledge gap or open problem, 2) a recipe for an experiment, 3) observations and 4) interpretations.

2.1 The knowledge gap

The knowledge gap or open problem is usually stated in the sections *Introduction* and *Related Work*. Start the introduction by saying what is already known about your topic. Ideally, you present a series of statements that are widely accepted and uncontroversial. Explain the solutions that have been put

forward to solve the open problem. Authors and scientific communities sometimes prefer to mention the known facts and well-established solutions in a separate section, *Related Work*. Having a separate section for related work highlights the fact that other people should be given credit for it and, conversely, that you alone should be given credit for the introduction.

As soon as you have established what others have found before you, identify the *gap*. Say what is currently unknown about the subject matter, what researchers before you have overlooked or been unable to explain. What problems cannot be solved by the current state-of-the-art approaches? Often, authors *justify* their work in the introductory paragraphs: why is this problem worth solving? What *could be* possible if only this problem were solved?

At the end of the introduction, showcase your "plan of attack". Explain how you will go about improving the current situation, that is, how you will bridge the gap or provide a better solution. Summarise the remainder of your work in a few words, only to be sure that readers will not go astray. Some people (the sort of people I will hereafter call "completionists") will expect that you explicitly mention which section will deal with which aspect of your work (as I have done in the introduction to this guide).

2.2 The recipe

The recipe for your experiment is an orderly collection of ingredients (conventionally called *Materials* or *Data*) and instructions (*Methods*). If this description reminds you of a cooking recipe, it is because the materials and methods sections are indeed very much like a cooking recipe. After all, different people following the same cooking recipe should arrive at the same results[1]. The most important role of the materials and methods sections is to ensure that your results can be *reproduced* by someone else. Reproducibility of results is the cornerstone of modern science and is absolutely central to your thesis![2]

2.2.1 Ingredients

Therefore, the materials section informs your readers where to find the ingredients of your experiment, in case they want to repeat it for themselves. And nobody will blame them for it, because many people have a hard time believing things they have not seen with their own eyes. If the materials can be obtained somewhere, indicate where they can be found (a link to a web address, for example). If, on the other hand, you have produced this material yourself, it is your responsibility to make it available to the scientific community. Depending on the nature of your data (a word I will use interchangeably with materials), other cannot possibly arrive at the same set of measurements or numbers. Publish the raw data or make sure there is a way to contact you, so that people can request the materials.

That is, people should be able to request your data within reason. If you forgive a short digression, a conservative individual once famously requested a scientist to hand out "the real data" that, he claimed, had not been published yet. Andrew Schlafly, an editor of Conservapedia and zealous disciple of Intelligent Design, demanded that Prof. Richard Lenski, a microbiologist, give the raw materials of his study to everyone. The conversation between them turned into the highly entertaining "Lenski affair" that you should read in its entirety[3]. Here is an excerpt of Prof. Lenski's second reply:

[1] In principle, at least. In reality, the outcomes will depend on people's cooking skills – and this is where the metaphor breaks down. If a researcher claims that a result cannot be reproduced, this is never taken to mean that they lack the skills to do it right.

[2] Unfortunately, reproducing other people's results is rare in Computational Linguistics and very often, authors do *not* give enough information. It is a very sad state of affairs and you, a young and promising researcher, should set a good example.

[3] See e.g. the RationalWiki article here: *http://rationalwiki.org/wiki/Lenski_affair*.

> It is my impression that you seem to think we have only paper and electronic records of having seen some unusual E. coli. [...] It's not that we claim to have glimpsed "a unicorn in the garden" – we have a whole population of them living in my lab! [...] So, will we share the bacteria? Of course we will, with competent scientists.

Obviously, microbial life can only be given to trained biologists because they are a serious health hazard. As a Computational Linguist, you are unlikely to work with bacteria and it is equally unlikely that you will be approached by a concerned Christian because of your published work.

Besides reproducibility, the materials section is there to simply *describe* your materials, so that you can develop the argument you are making in your thesis. So, describe the data and list important properties, facts and figures about it. For language resources in particular, those important properties are called *descriptive statistics*. The completionist would also recommend that you explain *why* you have chosen this kind of material and what are its advantages over the material you have not used.

2.2.2 Instructions

The instructions are sometimes not listed in the materials section, but separately in *Methods*. In there, you say what needs to be done to the data to transform it into your results.

First of all, explain the experimental setup. What kind of experiment did you conduct and what were the conditions? Again, if you are unsure whether to include a particular bit of information, think about reproducibility. Only include it if it is strictly necessary to reproduce the results. List all the equipment you have used and all existing tools you have plugged into your processing pipeline. Cite the authors of those tools to give them proper credit for writing useful software. If, during your experiment, the data is sampled in some way because of its size, name or explain the sampling method. Did you perhaps exclude some of the data during the experiment, e.g. discarding it because it was unreliable?

Introduce all the math that is necessary to understand your experiment. If statistical analysis is part of your work, list all the statistical tests you have used. Note that statistical testing is not a panacea to all problems and that there is serious doubt about its usefulness. I will not discuss the issues in detail (the interested reader is referred to Johnson, 1999 and Ziliak and McCloskey, 2008), but the important lesson is: it is not enough to report that your results are "statistically significant". You must by all means also report the sample size, the nature of the statistical test and the effect size. Then go on to explain how you calculated the results that you are going to present in a later section. What evaluation metrics did you use?

All of the above applies to you because you are an academic, but the following applies specifically to you because, as a Computational Linguist, you have a background in software engineering. If you write software that performs the steps you describe in the methods section, make the code available to others. Ideally, you place your code in the public domain with a license that allows unrestricted non-commercial use with attribution, for example on Github.

If you are confident that this leaves everyone with just enough information to be able to understand and reproduce your experiment, move on.

2.3 Observations

The observations you make during your experiment are listed in the *Results* section of your work. Relate to the readers what you have observed, after carrying out the instructions with the ingredients you have described earlier. Do not jump to conclusions in this section (or rather, do not even jump to

the *Discussion* section)! The results section is a sober, down-to-earth description of what happened. Try to be as objective as possible. You can only be truly objective if you report *numbers* instead of describing the results with imprecise words like "good" or "improvement".

The results and the discussion sections are sometimes collapsed into one big section that describes both – which is a bad idea, in my personal opinion. Authors that do this are prone to confusing facts with conjecture. Keeping the factual results and the speculative discussion in separate sections helps keeping them separate in your mind.

2.4 Interpretations

The next section, the *Discussion*, is about what the results *mean*, according to you. The results themselves, most likely numbers, are meaningless on their own and require an *interpretation*. If there is creative freedom in a scientific work, this is the section where it has its place. All of your readers will be aware that there are many interpretations of your results and that you cannot give a comprehensive list of them. Relate the results to the knowledge gap you have identified in the introductory sections. What do the results mean with respect to your overall goal, i.e. to bridge the knowledge gap or provide a new solution to an old problem?

Finally, you will write the very last section of your work, the *Conclusion*. It does not have its own section in this guide because, essentially, the conclusion section never presents novel information. Instead, it a summary of your main findings, the most central tenets that readers should bear in mind.

Now that we have established the contents of your thesis, we are left with discussing 1) when and in what order you should assemble the pieces (Chapter 3) and 2) how you should write (Chapter 4).

3 Writing Process

This section is dedicated to the *Process* of writing your thesis, which means, roughly, when to do what, in what order. Describing the genesis of a thesis in this way is a gross oversimplification of course, but it will get us started.

3.1 Order

The order in which you should write the sections of your paper differs significantly (no, I cannot give an effect size for this finding) from the order of publication. Having come thus far in your academic education, you are well aware of the format of scientific publications, and Chapter 2 of this guide also hints at this order. Figure 1 illustrates how the orders are different. In the left part of the figure, you see the stereotypical order of sections in published research. Yet, writing your thesis in exactly this order will be cumbersome – do not write the chapters in the order of publication.

The ordering of items in Figure 1 does not imply that you should only start writing the materials and methods section as soon as the bibliography is finished. It implies that you can start working on the bibliography earlier than anything else, without even knowing the materials and methods you are going to be using. The materials and methods section – the instruction manual for other people – takes precedence over many other sections because it can be written before the experiment starts. You can start writing up the results as soon as you have observed them because this section is concerned

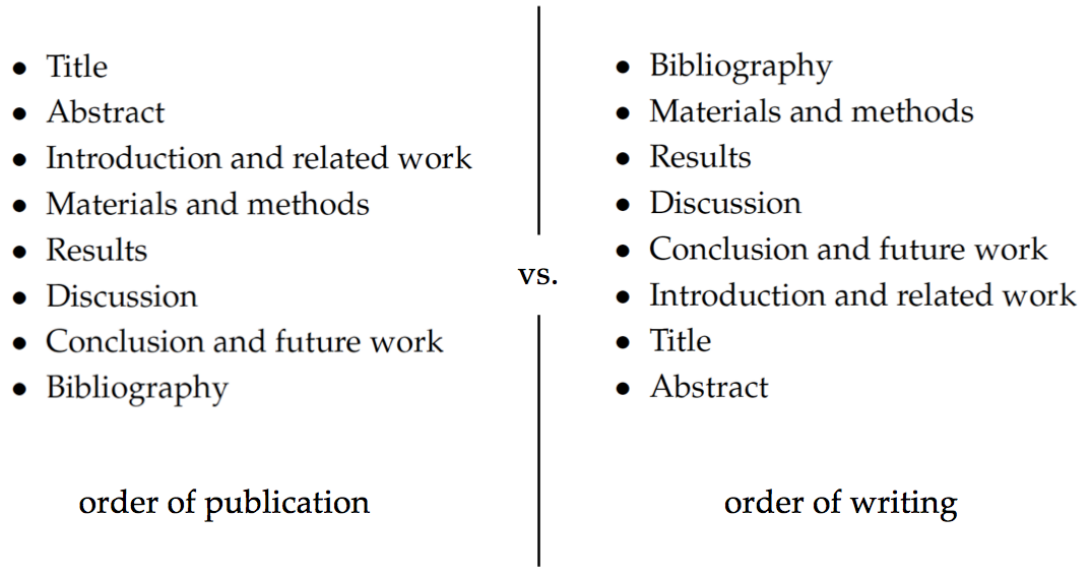

Figure 1: The order of sections of publication vs. writing

only with objective fact-gathering. The discussion naturally depends on the results, and the conclusion depends on both the results and the discussion.

Only after working on (by which I do not mean "having finished") all of those sections you will have enough insight and perspective to write the introduction and choose an accurate title. The abstract comes last because it effectively distils your whole thesis into a few key sentences that summarise everything.

Apart from the fact that you *can* start to write the materials and methods sections before conducting the experiment, there is also good reason why you *should* start early. Once your experiment has finished, you will be reluctant to redo everything, just because, while writing the methods section, it occurred to you that there is a serious flaw in the experimental setup. It might occur to you only once you start writing the methods section because writing down your abstract ideas requires that you think about them very seriously.

The most important message of this section is: start writing before the end of your experiment. The processes of writing and experimenting are mutually beneficial and this will have an impact on the quality of your work.

3.2 Writing strategies

This section discusses strategies that will help you organise your writing process. Collect information in an organised manner, start with the general picture and do not be reluctant to restructure your work.

3.2.1 Collect information

Start collecting and managing relevant information as early as possible. Keep digital notes about the data, the experiment and your observations. You can think of it as a lab diary, even though you might not actually be working in a laboratory. Ideas do not come to people's minds upon request – write them

down as soon as they occur to you.

Gather citations from the start and do not manage them by hand, use a tool instead. Even if you are not an early adopter of new technology, using software that stores your references is incredibly useful. Citation tools include BibTeX, Zotero, Citavi, Mendeley and Endnote.[4] Using such a software means that you will not have to write the bibliography yourself, it will be generated automatically for you. The references section of this guide was generated automatically, for instance.

3.2.2 Start broadly, zoom in later

There are several techniques to start broadly if you embark on a writing project. A useful first step could be to produce a mind map of what needs to be done. After that, you should definitely draw up a *skeletal outline* of your work. The skeletal outline has nothing more than the structure of the thesis and working titles for all the sections. At this point, it might be a good idea to show your outline to someone else. Tell them what each section will be about. If you cannot summarise a section in one single sentence, it is probably too broad.

Do not write sections and paragraphs in one go. Before you actually write them, represent each section with a single *topic sentence* that summarises the whole section. Then do the same for individual paragraphs. Topic sentences will be explained in more detail in Chapter 4.

3.2.3 Restructure your work

There is nothing wrong with restructuring parts of your work and you should not be afraid to do it. The mere fact that you invested time in writing something does not automatically make it correct or useful and there is always a way back.[5] Remember that most people are much more critical of other people's writing than of their own and act accordingly.

By now, it should be very clear to you what the content of a scientific thesis is – and it should be reasonably clear to you how the process of writing it could be organised. The next chapter will discuss the *style* of scientific writing.

4 Writing Style

This section leaves behind the topic of *what* to say and is about *how* to express your ideas. You do not only need to have ideas in the first place, you must also be able to communicate them effectively.

Above all else, strive for clarity and brevity, but not for one at the expense of the other. Every piece of writing is a tradeoff between the two, because the shortest document has zero length of course, but its content is rather unclear! "Complicated" is not the same as "scientific", even though, looking through German scientific texts, one must be forgiven for having this impression. Be clear and say what you mean – and your style of writing will still be regarded as scientific.

Why exactly should you be short? Because nobody likes to waste time reading stuff that is not important. This applies to all of your readers, not only your busy advisor. Obviously, what you can leave out is determined by your readers' level of knowledge. Picture your audience – what do they need to know? The next sections give helpful advice on how to be clear and short.

[4] The University of Zurich itself offers some of those tools and support for them: *http://www.id.uzh.ch/dl/sw/angebote/lit.html*.

[5] The *point of no return* is a popular theme in psychological theories about decision making. While there certainly are irreversible actions, the human psyche often tricks itself into thinking that there is no returning back, even though there would be (*http://psychology.wikia.com/wiki/Point_of_no_return*)

4.1 How to be clear

Writing clearly has nothing to do with intuition. Rather, there are a number of tangible principles that bring clarity to your writing. Use topic sentences and never mention more than 1 point per paragraph. Pay attention to word order, exploit parallel structures and transitions between sentences. As the last polish, make sure that there are no typos in your work and that the grammar is impeccable.

4.1.1 Topic sentences

Make the first sentence of each paragraph a *topic sentence*. A topic sentence summarises the whole paragraph and is a very eloquent way of telling your readers what they can expect from the paragraph. Compare a (fictitious) paragraph with a topic sentence, where the main statement is at the beginning:

> **Reordering the source text is an improvement to our SMT system that yields +1.5 BLEU points.** We have introduced a reordering component as part of the preprocessing pipeline because manual evaluation of the SMT output suggested that many errors are due to word order. Since our system translates from German to French, the word order might indeed be an obstacle. For instance, German past participles are placed at the very end of sentences, but in French are part of the main verb cluster.

To a paragraph where the main statement is at the end:

> We have introduced a reordering component as part of the preprocessing pipeline because manual evaluation of the SMT output suggested that many errors are due to word order. Since our system translates from German to French, the word order might indeed be an obstacle. For instance, German past participles are placed at the very end of sentences, but in French are part of the main verb cluster. **Reordering the source text is an improvement to our SMT system that yields +1.5 BLEU points.**

4.1.2 One point per paragraph

The paragraph should be your unit of information and every paragraph should only be about one point. If you cannot decide on a single "point" for your paragraph, split the content into at least two paragraphs. In general, the average paragraph in an academic publication is too long – longer than the reader's span of attention. On the other hand, if your paragraph does not have anything to say on its own, combine it with another paragraph.

4.1.3 Word order

Maintaining an SVO (subject verb object) word order where possible makes your sentences more clear. To English and German brains, SVO order is exactly how their thinking progresses (this is a wild speculation that I do not have proof for). Rather than putting in front one of the objects as in the following sentence that could have been part of Section 2.2.1 of this guide:

> On how to find the ingredients you will write in the *Data* section.

Start with the subject and verb, then proceed with the objects:

> You will write on how to find the ingredients in the *Data* section.

4.1.4 Parallel structures

Parallel structures are junks of writing that are deliberately repetitive. Repeating structures on purpose ties the text together.[6] Here is an example from Section 2.3 of this guide:

> **Keeping** the factual results and the speculative discussion in **separate** sections helps **keeping** them **separate** in your mind.

Where two terms are deliberately repetitive, "keeping" and "separate". The repetition establishes a very close relationship between the clauses. Here is another example from Section 2.2.1 where the parallelism spans two paragraphs:

> Publish the raw data or make sure there is a way to contact you, so that **people can request the materials**.
>
> That is, **people should be able to request your data** within reason.

4.1.5 Transitions

Transitions help to maintain flow as you move from one sentence to the next. Put another way, a transition is a device that shows how sentences are related. The easiest kind of transition are transition words like *but*, *and*, *also*, *therefore* or *nevertheless*. They all make *explicit* the relationship between sentences. *Moreover* is a transition word you should use sparingly – it is somewhat overused in English writing. The following sentence taken from Section 3.1 makes use of transition words:

> **Apart from the fact that** you *can* start to write the materials and methods sections before conducting the experiment, there is also good reason why you *should* start early.

Apart from words that have this specific function, you can also transition between sentences by simply mentioning things *again*. Either use the exact same topic word again later in the text or paraphrase it very closely. Another helpful technique is to make a topic the object of the first sentence, and use it as the subject of the next, as I have done in the following sentences taken from Chapter 1:

> Writing scientifically is not just a matter of finding the right words that will, hopefully, persuade your **audience**. An **audience** of scientists can only be persuaded with facts that result from a well-thought-out experiment.

Or in Section 4.1.1 where I explain topic sentences:

> Make the first sentence of each paragraph **a *topic sentence***. **A topic sentence** summarises the whole paragraph and is a very eloquent way of telling your readers what they can expect from the paragraph.

4.1.6 A note on grammar

Writing correct English is paramount, because typos and bad grammar distract your readers from the content. Check the spelling as you write and let a native speaker of English review your work. Give the reviewer enough time to check your thesis, which is only possible if you finish writing it way ahead of the hand-in deadline.

[6] In fact, repetition is a very common rhetorical device. It is quite likely the most frequent type of rhetorical figure, because it immediately appeals to even the layman.

Pay special attention to tenses. Always use past tense for observations that you or others have made and use present tense for "generalities". Generalities are facts that are either obvious or common knowledge – things you definitely were not the first to stumble upon.

4.2 How to be short

Make your writing take up as little space as possible. Stay on-topic, avoid fog and fillers and do not evade the weaknesses of your work.

4.2.1 Stay on-topic

Do not say anything that has no bearing on the problem and topic of your thesis. This might seem obvious to you, but if you think about it, people say and write irrelevant things all the time. Specifically, do not mention something just to 1) make your readers aware that you know this or 2) to extend your list of cited works!

4.2.2 Fog and fillers

Leave out all the fog and fluff in your writing. A word is "foggy" if it blurs the picture of what you are trying to say. A fog word is ambiguous or vague. Fluff words are fillers that have no meaning on their own and are interspersed in texts for no good reason. That is, the sentences appear to be longer and the total word count will increase significantly - but that's about the only advantage of fluff words. Both fog and fluff are in the way when readers are trying to understand your work and will delay them considerably.

4.3 Word games

Do not hide behind word games and euphemisms. Like fog and fluff, word games confuse your readers. Also, it puts their patience to the test because they will find out eventually. Science is not advertising, there is no need to gloss over the weaknesses of your work. Quite on the contrary, mentioning the limitations of your work is regarded as good scientific practice.

5 More general food for thought

Nowadays, native speakers of English are at an advantage because English is the lingua franca of pretty much every scientific community in the world. So, seriously consider having your work proofread by a native speaker. This situation might change in the future, though – and if it does, English speakers will realise that it's pretty hard to write scientific articles in a foreign language. Meanwhile, do not despair and practice writing scientific English as much as you can.

On a related note, it is a common misconception in the German-speaking world that the use of impersonal constructions should be encouraged. For example, people would advocate writing "Problem X was investigated" instead of "We have investigated problem X". The problem with the former is that there is no way of knowing you has done what. You and your colleagues? Authors before you? Your debugging tool? Always make it clear who is taking the responsibility.

Always value content over form. Here is an ingenious trick to see for yourself whether the content has merit on its own, independent of how it is presented. Typeset the content in a font that you don't like and see whether it still makes sense[7]. If it stops making sense it was probably the font that made you like it.

6 Topics that this guide does not cover (yet)

This is by no means a comprehensive guide to writing a scientific thesis. In the future, I (or anyone else, for that matter) could consider extending it and also talk about

- **how to find a research topic in the first place:** since finding a suitable research question makes writing your thesis a lot easier
- **the role of the advisor:** expertise they need, how much advice you can expect, is the advisor also a proofreader?
- **scheduling work:** how to finish writing in time, planning ahead
- **how to find resources:** libraries, scholarly collections, databases, conference indexes
- **tables and figures:** content, layout, axes, scales, captions, colours, size
- **thesis formatting:** working with Word or LaTeX, developing a coherent style, fonts and sizes, indentation, aesthetics
- **how to get your work published:** finding a publisher or suitable journal, submission process, peer-review

[7] See *https://www.reddit.com/r/LifeProTips/comments/318ilc/lpt_when_reviewing_something_youve_written_change/*.

References

Johnson, D. H. (1999). The insignificance of statistical significance testing. *The journal of wildlife management*, pages 763–772.

Katz, M. J. (2009). *From research to manuscript: a guide to scientific writing*. Springer Science & Business Media, Berlin.

Peat, J. (2002). *Scientific Writing: Easy When You Know How*. BMJ Books, London.

Ziliak, S. T. and McCloskey, D. N. (2008). *The cult of statistical significance*. University of Michigan Press, Ann Arbor.