

SoNaR Acquisition Manual

version 1.0

LT3 Technical Report – LT3 10-02

Orphée De Clercq* and Martin Reynaert †

(*)LT3 – Language and Translation Technology Team
Faculty of Translation Studies
University College Ghent
orphee.declercq@hogent.be

(†) Induction of Linguistic Knowledge Research Group
Tilburg Centre for Cognition and Communication
Tilburg University
reynaert@uvt.nl

URL: <http://veto.hogent.be/lt3>¹

October 25, 2010

¹The reports of the LT3 Technical Report Series (ISSN 2032-9717) are available from http://veto.hogent.be/lt3/publications_en.html All rights reserved. LT3, Faculty of Translation Studies, University College Ghent, Belgium.

Contents

1	Introduction	1
2	Contact	3
2.1	Identifying Text Providers	3
2.2	Contacting the Right Person	3
2.3	Negotiating	4
3	Basic principles	5
4	Possible situations	7
4.1	Text provider wishes more information	7
4.2	Text provider sends email permission	10
4.3	Text provider questions license agreement	10
4.4	Text provider does not answer	13
4.5	Unexpected developments	14
4.6	Statistics	15
5	Conclusion	16
A	Standard Agreement	17
B	Agreement for Publishers	19

Chapter 1

Introduction

Although strongly needed, there exists so far no standardized procedure for dealing with data collection and copyright negotiations, two crucial steps in corpus building. Thanks to the STEVIN-programme ¹ several new Dutch corpora, or corpora containing at least one Dutch component, have emerged. An essential prerequisite for all these corpus compilation projects was that they have to be completely copyright-cleared so as to ensure their widespread availability for both research and commercial purposes. This led to a semi-standardized procedure of dealing with copyrights and data collection in the Netherlands and Flanders which might prove useful for future corpus compilers, both in the Netherlands and abroad.

This document describes the nuts and bolts of what it takes to construct a reference corpus of written Dutch such as SoNaR. SoNaR is to contain 500 million words spread over more than 35 text types, including both traditional and new media texts. For more information about the SoNaR project we refer to Reynaert et al. (2010). All the text material included in the SoNaR corpus is copyright-cleared. For the traditional text types we were able to learn from the experience gained during other STEVIN-funded projects such as the Dutch Parallel Corpus ((Trushkina, Macken, and Paulussen, 2008) and (Rura, Vandeweghe, and Perez, 2008)) and the D-Coi project (Oostdijk et al., 2008), but new media text types such as blogs, chats and SMS present us with new data collection challenges. The principles described in this manual are all based on hands-on experience.

We would like to stress that we are not lawyers, merely (computational) linguists who were asked to compile a corpus that is copyright-cleared. So when it comes to discussing copyrights, we are only able to mention the difficulties we encounter and how these are surmounted. The license agreements we use were drawn up by the lawyers of the HLT Agency ². If you want to draw up your own license agreement we strongly advise you to consult a lawyer or legal adviser as the situation in your country may be very different.

Since there is no standardized procedure for dealing with text acquisition and IPR-settlement, we try to formulate some basic principles. As a consequence, this is not an exhaustive manual, and many more questions or other difficulties might arise. If you have been assigned the task to collect data material for a specific research project and you have absolutely no idea where to start, this manual might nevertheless just be the ideal starting point.

In the following sections we describe the acquisition process in great detail ranging from contact-

¹A research programme funded by the Nederlandse Taalunie (Dutch Language Union) whose purpose it is to stimulate the development of language and speech technology in Flanders and the Netherlands, for more information please consult www.stevin-tst.org

²This is the Flemish-Dutch Human Language Technology Agency, see www.inl.nl/en/tst-centrale

ing the right person (Chapter 2) and laying down some basic principles (Chapter 3) to discussing the different, sometimes difficult, situations that may arise (Chapter 4). Everything is illustrated with example letters/emails, which were originally written in Dutch, but are here translated into English. We end with some concluding remarks and some prospects for future work (Chapter 5).

Chapter 2

Contact

All texts included in the corpus have to be in electronic form. This explains why the first place to look for text material is on-line. Although there exist many websites wide open to the public, it is a misconception to believe that copyright laws do not apply to these websites. For each piece of text you wish to include in your corpus, copyright clearance has to be sought, unless the website specifically mentions a particular license which obviates you from doing so.

2.1 Identifying Text Providers

If you combine your world knowledge with adequate queries in Google you are able to find text providers for almost every text type that will be included in your corpus according to its typology.¹

A data collector should decide beforehand which text type he/she is going to look for. Legal texts, e-magazines as well as chat boxes will probably not be found on the same website. Each text type requires its particular queries and other contact persons: profit versus non-profit, organisations versus individuals, teenagers versus elderly people.

Deciding whether a text provider is valuable or not is rather subjective. Still, some objective criteria can be discerned to validate a possible text provider.

- Is this a large company?
- Is it a public institution?
- How many Google hits does the site get?
- What amount of text material does it hold?

2.2 Contacting the Right Person

Finding a possible text provider is one thing, but contacting the right person is another. Your contact person should be someone who might be able to help you with your request immediately

¹In this manual we do not discuss corpus design, typology, ... There already exist many reference books treating this aspect. For example: Kennedy (1998) and Lüdeling and Kytö (eds.)(2008).

or at least able to bring you into contact with the right person(s).

Nowadays, most organisations (both profit and non-profit) have someone in charge of communications. Other possible contacts might be the webmaster, the legal department or the management. This information is most often mentioned on company websites: for example at the bottom of a press release, in the disclaimer, in the sections with contact information, corporate governance or company structure. Always try to find personal email addresses.

There are, however, also organisations where you will only be able to click on a certain 'contact' button or on a box where you can type in your question. In these cases, you should be aware that you are contacting someone who might be able to help you find the right person.

2.3 Negotiating

Negotiating is the most important part of the entire acquisition process. Different situations may arise and some things have to be repeated over and over again. The ultimate goal of this negotiation process is to get permission in writing from the text provider that his/her text material can be included in the corpus.

For SoNaR two types of license agreements were developed by the Dutch HLT Agency: one standard agreement (Appendix A) and one agreement for publishers (Appendix B). Our aim is that the negotiation process should always end with the text provider signing three copies of the license agreement. Subsequently, these copies are sent to the other parties, who each sign their part. Finally, every party receives a completely signed copy of the agreement.

Ideally, the perfect acquisition process consists of three or four steps. First, you send a general letter to the text provider, he/she answers that he/she is interested and would like some more information. You reply, give some more information and explain that you work with a standard agreement which has to be completed. The text provider agrees with this agreement, completes and signs it, and sends it back to you.

In reality, however, the actual negotiation process is far more complicated. In Chapter 4 the various situations are discussed and documented with examples. These are all real-life examples of situations that occurred during the compilation of SoNaR.

Chapter 3

Basic principles

Before we can start to discuss different situations that might arise and how to deal with these, there are some basic principles that always have to be followed when contacting a possible text provider: be polite, to the point and always remain friendly.

Know your project

It is very important to know your project inside out, before contacting a potential text provider for the first time. Make sure you will be able to answer every possible question and anticipate all sorts of situations.

One very important issue you ought to consider is how to protect the text material in the corpus. This is a first stumbling block for practically every text provider, who all will want to be reassured. In most cases text providers do not want you to alter any text material at all, but on the other hand they might also demand particular data to be anonymized. They will insist on acknowledgement and most importantly they will want to be reassured that you will not distribute the corpus to third parties without taking proper measures to safeguard the text owners' rights. Therefore it is necessary to anticipate how the eventually realized corpus will be distributed.

Here are some issues you have to consider:

- Will the end user be able to see full text or only snippets?
- Will the end user be able to download anything?
- Will the end user have to sign an agreement to be able to use the corpus?
- How will you ensure that the end user will not further distribute the material to third parties?

First Contact

Once you have thought about all this, the actual negotiations can start. The first contact is very important. The next example shows the formal letter/email we send to all the prospected text providers of the SoNaR project.

This first contact is actually a sort of request letter in which you have to give the reader enough background information and appeal to your reader to do something for you. It should at least contain the following parts: salutation (if you have the name of your contact person: use it),

request, background information, repetition of the request and use of an anticipating closing formula.

Dear Sir X or Dear Madam X,

We are writing to request your permission to include Dutch text material from X (e.g. the website www.hogent.be) in the SoNaR corpus.

As a partner of an academic consortium including the University College Ghent, KU Leuven, Tilburg University and the Universities of Nijmegen, Twente and Utrecht, we are working on the compilation of a large corpus of written Dutch containing no less than 500 million words. The objective of SoNaR is to create a reference corpus of written Dutch as it is being used in the Dutch speaking area (the Netherlands and Flanders). Both language variations will be represented equally in the corpus, which is unique. For more information about the project please consult the SoNaR website: <http://lands.let.ru.nl/projects/SoNaR/>

The SoNaR-project is carried out within the framework of the STEVIN-programme of the Dutch Language Union. STEVIN stands for Speech and Language Technological Essential Data and Facilities in Dutch. It is a long-term programme, funded by the Dutch and Flemish governments, whose purpose it is to stimulate the development of language and speech technology in Flanders and the Netherlands in order to consolidate the position of Dutch in the modern information and communication society. More information can be found on the leaflet attached to this mail and on <http://taalunieversum.org/taal/technologie/stevin/english/>.

We are looking for electronic text material and we would like the corpus to contain as great a variety of different text types as possible. We are for example looking for newspaper articles, press releases and brochures, but also text material coming from newer media such as websites, chat sessions and text messages will be included.

This is why we are contacting you, as X (e.g. webmaster of the website www.hogent.be) you might just be the ideal person to help us include text material in the corpus. For example the text material X, Y and Z (e.g. written under info, press releases and ...) could be extremely useful for our project.

Therefore, we would like to ask your permission to include text material in the SoNaR corpus. During further contacts we can negotiate permission clearance and copyrights. The corpus will be used for research purposes, such as translation technology, translation studies, linguistics and language didactics.

Should you have any further questions, please do not hesitate to contact me.

We look forward to hearing from you,

Yours sincerely,

Contact Box

If you contact someone in a more indirect way, via the contact button or question box (see 2.2), you should not fully explain the project. Briefly introduce yourself and your project and ask for a person who might be able to help you. This might look like this:

Dear Sir or Madam,

I am working as a research assistant at the University College Ghent. Currently we are involved in a project that aims to collect as much written Dutch text material as possible.

The text used on X (e.g. www.hogent.be) could be very useful for this project. Who can I contact to give further information and perhaps to obtain permission?

Thanks in advance and kind regards,

If this request via the contact box is successful, you will get a contact address after which you can contact the appropriate using the first contact example.

Now that we know how to establish a first contact, we can take a look at the next chapter in which different situations and problems are introduced, together with their solutions.

Chapter 4

Possible situations

This chapter is a summary of a number of different situations that were encountered when negotiating acquisition and copyright clearance for the SoNaR project. This is not an exhaustive list and other situations might arise, but the examples below illustrate which questions and problems always return.

4.1 Text provider wishes more information

If the text provider asks for more information, this is good news: it generally means that he/she is interested in participating.

Some examples:

Dear

I have received your request to use text material from X for research purposes. Could you please give me a call to discuss your request in closer detail?

Kind regards

Dear X

You can contact me for all the text material on X. Considering your request I believe it might be useful to meet in person.

Look forward to hearing from you.

Kind regards

In theory I don't have problems with your request, since X is a public website. How do we proceed? How do you decide which texts are useful to include in the corpus, how do you download the text material, ...

Dear Ms X,

First and foremost my apologies for this late reply.

Basically, we are willing to cooperate and we might allow you to include text material. Could you please inform me how this cooperation will proceed?

Hello,

You can count me in. Just let me know what the next steps are.

Best wishes,

Dear Ms X,

Could you please give me a call? I have some further questions?

From these examples we can derive that there are two kinds of respondents: those who want to continue negotiating via mail and those who prefer a more personal approach (by telephone or in person).

These examples also nicely illustrate which are the first questions possible text providers might ask.

Answer by mail

The examples above reveal that most text providers want to know more about the practical side of the collaboration, which is a perfect occasion to introduce the license agreement. Also mention that this collaboration will not take up much time or work.

Some examples:

Dear Sir X,

Thank you very much for your reply and for your interest in the project.

A prerequisite for text material to be included in the SoNaR-project is that it has to be copyright-cleared, which is why we seek permission clearance.

On your website I did not find information concerning copyrights. Am I correct when I state that everyone who writes text on X is aware that his/her text is in the public domain or has transferred his/her rights to your site? If this is not the case I would like to make sure one way or the other that everyone agrees to his/her text material being included in the corpus.

In order to arrange permission clearance we work with the standard license agreement, attached to this email so as to allow you to get acquainted with its terms and conditions. I trust you will let me know whether you can agree to these.

For the actual data transfer we are thinking about the online archive. Once we have received your permission, our computational linguist will be able to download all the material.

Furthermore, I would also like to mention that we do not alter any text material. The only thing we can do is add metadata (name of the author, text type, source,...) and run our language technology tools on it (we have for example a tool that adds morpho-syntactic information to every word).

I hope you now have a clear image of the project. Please do not hesitate to contact me if you have any further questions.

Thank you for your time and kind regards,

Dear Ms X,

Thank you very much for your reply. Wonderful news that you are willing to cooperate.

To be able to include text material in the corpus our commissioner, the Dutch Language Union, requires permission in writing. We work with the standard license agreement, attached to this email so as to allow you to get acquainted with its terms and conditions. I trust you will let me know whether you can agree to these.

The actual data transfer will not take up much time. Once we have received your permission my colleague will download all the text material.

In addition, I would like to mention that we do not alter any text material. The only thing we can do is add metadata (source, author, text types,...) and run our language technology tools on it (we have for example a tool that adds morpho-syntactic information to each word).

Dear Sir X,

Thank you very much for your mail. Wonderful news that you wish to cooperate.

To be able to include text material in the corpus our commissioner, the Dutch Language Union, requires permission in writing.

We work with the standard license agreement, attached to this email so as to allow you to get acquainted with its terms and conditions. I trust you will let me know whether you can agree to these. If you can agree to its terms I would like to ask you to fill in the agreement, print it three times and to sign the three copies.

Please send these three copies to X.

Answer in person

First make an appointment when you will call or come by.

Dear Sir X,

Thank you very much for your reply and interest. Is it convenient if I give you a call tomorrow morning?

Kind regards,

Dear X

Wonderful news. I can thus assume that X wishes to cooperate to the SoNaR project.

I will gladly give you some more information about the project in person. Please suggest a date and time. Look forward to meeting you.

When talking in person, make sure that you are adequately informed about the activities of the company/organisation. Take a closer look at their website and make sure that you mention the right names of text material (magazine, newsletter, ...), persons, products and so on.

Although conversations cannot be predicted, it is possible to anticipate some probing questions text providers might ask.

- What is the project about?
- Why do you want our text material?
- What is a corpus?
- Who will use the texts?
- Will the texts be altered?
- What do I have to do?
- Why do I have to sign a license agreement?
- What about the commercialisation of the project?

- What is in it for me?

Once you have provided an answer to any or all of these questions, the conversation will advance considerably.

Meeting text providers in person is rather exceptional, first consider whether the text provider is important enough to make the extra effort. There is a difference between meeting a journalist from a local newspaper and meeting the editor-in-chief of a newspaper concern.

Also bear in mind that the license agreement will probably not actually be signed during this meeting. The negotiating process may still take a long time, afterwards.

4.2 Text provider sends email permission

This is good news; it means that the text provider wishes to participate. Explain that you need a permission in writing. An example of this with a possible reply is:

Dear X,

Since our website is openly available to the public you can freely use our text material for the project.

Kind regards,

Dear X,

Thank you very much for your reply. Wonderful news that you are willing to cooperate to SoNaR. In your last reply you already gave a permission. Our commissioner the Dutch Language Union, however, requires permission in writing.

We work with the standard license agreement attached to this email so as to allow you to get acquainted with its terms and conditions. I trust you will let me know whether you can agree to these.

Thanks in advance for your time and kind regards.

4.3 Text provider questions license agreement

Negotiating IPR (Intellectual Property Rights) agreements is probably the most difficult part of the entire acquisition process. It may take a long time and frustrating situations may arise. Remember to always remain calm and polite, you are after all asking for a favour for which basically no compensation is given¹.

For the SoNaR project we worked with two agreements: one basic agreement (Appendix A) and one for publishers (Appendix B).

The major stumbling block during copyright negotiations is the commercial aspect of our agreements. There is a small section in which it is stated that all text material can be used for both research and commercial purposes (Article 1.2 in the agreement). Then there also arise situations where the text provider wishes to adapt the agreement or refuses to sign. We next illustrate these problems with examples.

¹or very little, e.g. in the form of some publicity. SoNaR has no funds for buying any rights.

Commercial aspect

When a text provider is confronted with the IPR-agreement for a first time, his initial reaction may be to disagree because of its commercial aspect.

In this situation it is important to explain that for SoNaR and for other corpora the notion 'commercial' is different than he/she assumes.

Some examples:

Dear X,

Thank you for your interest in our website, it is always pleasant when someone shows interest in our ins and outs.

We have no problems that you use our text material for a research project. That is, provided on some conditions:

- You (and any partners) will not pursue commercial purposes
- Also in the communication following the project no commercial purpose can be foreseen

I hope this is clear to you, otherwise we will hear so.

Greetz

Dear Sir X,

Thank you very much for your reply and for your interest in the project. Our actual goal is to develop a reference corpus of written Dutch.

It is indeed so that the corpus will primarily be employed for research purposes, the end product can nevertheless also have a commercial aspect. Please let me give you some more information.

The general conditions on your website state that every user is aware that the text he/she publishes is in the public domain and that he/she gives away his/her rights. This statement is extremely valuable for us because it implies that when we reach an agreement with you we are also able to reach an agreement with every user.

Should we receive your permission to include your text material in our corpus (we are thinking about, among others, the archive,...), you would help us very much. I would like to stress that we do not alter any data, the only thing we can do is add metadata (source, author,...) and run our language technology tools on it (e.g. we have a tool that adds morpho-syntactic information to every single word).

What we need is a permission to distribute the text material for research purposes. The corpus will be used for language technology and research, language didactics and applied linguistics.

The commercial aspect is only included if the eventually processed corpus is to be employed by a commercial company working with language and speech technology. But then it is extremely important to realize that the end user will never be able to download the material as a whole. In a corpus, it is not the content that matters, but words and word forms that are being used, we only work with derivatives of text.

I hope this information has given you a clear image of the corpus. We work with the standard license agreement attached to this email so as to allow you to get acquainted with its terms and conditions. I trust you will let me know whether you can agree to these.

Thank you very much for your time and kind regards,

Ms. X,

We are willing to give text material for your project but only if it is used for research purposes. Commercial purposes will have to be excluded.

Dear Sir X,

Thank you very much for your reply and for your interest in the project. The SoNaR corpus will indeed be primarily used for research purposes. There is, however, also a commercial aspect included in the agreement.

That is why we work with the standard license agreement, which is attached to this email so as to allow you to get acquainted with its terms and conditions. I trust you will let me know whether you can agree to these.

Please do not hesitate to contact me if you have any further questions.

Adaptation

It is possible to adapt the agreement. This adaptation, however, has to compensate for the efforts that are being put in the corpus creation. It is thus a question of wheeling and dealing between both parties.

Possible adaptations are:

- Text provider wishes everything to be anonymized;
- Text provider cannot allow commercial purposes;
- Text provider wishes to add some extra stipulations.

We do not have an example of this because of privacy reasons, but this does occur. For example, if the text provider cannot allow use for commercial purposes, we first explain more carefully that commercial use does not entail republication of any of the texts and thereby try to nevertheless persuade them. If this is not possible we are able to agree to this term, provided that an extra stipulation is added to the agreement stating that we are allowed to contact the text provider again in the case a user of the corpus would ever wish to be able to use the provider's specific subpart of the corpus for commercial purposes.

Refusal

Even if you have put a lot of effort in the negotiating process, it is still possible that the text provider refuses to sign the agreement in the end. Even then it is best to still thank them for their time and cooperation.

An example

Dear,

You are free to use the text material on our website for your project but we find signing an agreement exaggerated.

Good luck and best wishes,

Dear Sir X,

Thank you very much for your reply.

I can understand that you find the agreement a bit exaggerated, but I am afraid this is the standard procedure of the Dutch Language Union. All the material we get at our disposal for a research project has to be copyright-cleared to avoid future problems.

Thanks again for your time and cooperation.

Yours sincerely,

4.4 Text provider does not answer

If you send a first email to a contact person make sure that you add a 'read certification', then you know when your mail has been read.

If you have not received an answer within two - three weeks, it might be possible that your mail was lost in the mailbox or that the text provider did read your email but has forgotten all about it. A third possibility is that he/she is just not interested.

For a text provider, participating in a research project is not a core activity, it is something that comes second or third place, so you may have to gently remind him/her from time to time.

After sending the IPR agreements, it is also common not to receive a reply.

Some examples:

Dear X

About a month ago I contacted you with a request to include text material from X in the SoNaR corpus. The text material on X would be extremely useful for our research project.

In the Copyright statement I read that I should contact the editorial office, but I have not received a reply since.

Have you received my mail correctly and maybe already found the time to discuss this?

Thank you very much for your time and best wishes,

Hello,

We apologize, we completely lost track of your request. I will ask around and let you know something more asap.

Kind regards,

Dear X,

About two weeks ago I sent you a license agreement for the SoNaR project.

I was wondering whether you received this correctly? We hope that X is still willing to cooperate to SoNaR, the text material on X would help us considerably.

I would like to ensure you again that we do not change the texts and that the end user will never be able to see or download the text material as full text.

Thank you very much for your time and best wishes,

Dear,

A signed copy of the agreement is lying on my desk. Could you please give me your fax number, then I'll fax it right away.

Yours sincerely,

Dear X

In December we met to talk about the SoNaR project (a reference corpus of written Dutch representing the Dutch speaking area)

We are still extremely interested in including text material from X in our corpus.

Mid December I sent you two standard license agreements (a normal one and one for publishers). I was wondering if you received these and whether you have found the time to get acquainted with their terms and conditions? I attach the agreements a second time, just in case.

Hi X

My apologies that this is taking so long. I have repeatedly contacted my supervisor and forwarded your request. I hope this will be fixed shortly. Please keep me up to date.

Best wishes,

4.5 Unexpected developments

So far the situations described are those one intuitively expects, there were however also situations that we did not anticipate. These are listed below. Sometimes the text provider wants something in return or he/she has not filled in the agreement correctly or has sent back only one copy of the agreement.

Dear X,

I have just received the agreement, for which I thank you very much. There is a small problem, however, as we need three signed copies of the agreement: one for each party. My apologies, I should have made this clearer in my last email.

I also noticed that the name of the second person is placed on the wrong line. Please find attached a correct copy of the agreement.

Please, kindly sign the agreement a second time and send all three copies back to me?

Thank you very much and best wishes,

Dear X,

Thank you very much for your reply. I had already surmised that because chat conversations are more like private conversations we need permission from all the users.

We were also thinking about creating a chat box where the users first have to agree to the general terms of use, in which it would then be stated that everything they write will be included in the SoNaR corpus. Do you believe this is feasible and would you be able to help us?

Thank you very much for your cooperation and kind regards,

X,

Currently we don't have the time nor the resources to launch something like that, we have a lot of other work.

I suppose you don't have a budget to build/create a chat box?

Another unexpected development is that the text provider has already signed an agreement for a similar project with the Dutch Language Union in the past and that he/she does not wish to repeat this process.

Dear X,

Exactly one year ago we signed a similar agreement for another project also commissioned by the Dutch Language Union. We believe it is superfluous to repeat this process.

Dear,

I understand, I would like to ask you to send me an email message in which you state that you allow the people of SoNaR to use the same text material that was used for the other project.

The mail could look something like this:

X, represented by name, function, hereby gives permission to also apply the license that was used for the Y project to the SoNaR project.

This agreement includes all the text material that is presented on the website Z.

I hope you can agree with this. Thank you very much for your time.

Best wishes,

As can be derived from this example, an email with permission can only be accepted by the Dutch Language Union in exceptional cases. You should always try to draw up a new agreement for every new project.

4.6 Statistics

In conclusion, some figures are presented for an acquisition period of three months.

During this period, 65 organizations were contacted to participate in the SoNaR project: 33 organizations answered and 32 did not.

Within this time span of three months, we were able to conclude eight agreements. Nineteen text providers were interested but still had to discuss the agreement internally. Six others asked for more information but did not further reply after they had received the agreement.

This means that of all the possible text providers you contact, about 50% is interested and from these interested ones, about 25% is willing to sign the agreement.

This means that slightly more than 10% of the contacts you make end in success. Of course, the more organizations you contact, the more agreements you will be able to conclude.

Chapter 5

Conclusion

In this manual, we have tried to describe the nuts and bolts of what it takes to negotiate text data collection and IPR-settlements.

Everything ranging from prospecting and establishing the first contact to possible difficulties that might arise were discussed and illustrated with various real-life examples.

This manual is not an exhaustive list of situations and there will undoubtedly arise other situations during your personal negotiations. Nevertheless, this manual is a good starting point for everyone who wishes to create a corpus, collect text material and settle IPR-issues.

Future Work

This manual will be continually updated. Among other things, we are currently working on a separate chapter dealing with the new media text types, for which sending back and forth three signed copies of a license is simply infeasible.

For further information about the SoNaR project, please refer to the SoNaR-website at <http://lands.let.ru.nl/projects/SoNaR>

Appendix A

Standard Agreement

License Agreement

[Name of the company/organization/person], hereby lawfully represented by [Name, Function]; hereinafter named 'License Holder';
and

1. The Nederlandse TaalUnie (Dutch Language Union), a treaty organization having legal personality, situated and having its seat in the Hague, hereinafter termed 'NTU', hereby lawfully represented by Jeannine Beeken, director of the INL (Institute of Dutch Lexicology);
2. Radboud University Nijmegen, Faculty of Arts, hereby lawfully represented by Prof Dr Nelleke Oostdijk, hereinafter termed the 'Project Supervisor';

Both parties are at the same time collectively termed the Licensees

Article 1 License

1. The License Holder hereby grants to the NTU and the NTU accepts the right to use [Description of the Text Material], hereinafter termed 'Language Data' or have them used, to process them, make additions to them, to modify them, to retrieve them or to (re)use them for the construction of the SoNaR corpus, hereinafter termed 'Corpus'
2. The License Holder hereby also grants to the NTU and the NTU accepts, the right to multiply (a part of) the Language Data, to publish them, to reuse and utilise them, including a) the right to grant sublicenses to third parties for the use of the Corpus in research and education b) the right to include and/or integrate the Corpus in other systems such as corpora and corpus query systems and c) the right to grant sublicenses to third parties for the use of the Corpus for application development and the use of these applications for commercial and non-commercial purposes.
3. The NTU shall only grant a right to third parties to use the Corpus for the benefit of product development for commercial purposes under the condition that the Products shall not be recognisable in the new products to be developed by these third parties. The NTU shall not give the right to third parties to publish full texts or fragments of the Language Data.

Article 2 Disputes and applicable law

Appendix B

Agreement for Publishers

License Agreement for Publishers

[Name of the company/organization/person], hereby lawfully represented by [Name, Function]; hereinafter named 'License Holder';
and

1. The Dutch Language Union, a treaty organization having legal personality, situated and having its seat in the Hague, hereinafter termed 'NTU', hereby lawfully represented by Jeannine Beeken, director of the INL (Institute of Dutch Lexicology);
2. Radboud University Nijmegen, Faculty of Arts, hereby lawfully represented by Prof Dr Nelleke Oostdijk, hereinafter termed the 'Project Supervisor';

Both parties are at the same time collectively termed the Licensees

Article 1 License

1. The License Holder hereby grants to the NTU and the NTU accepts the right to use [Description of the Text Material], hereinafter termed 'Language Data' or have them used, to process them, make additions to them, to modify them, to retrieve them or to (re)use them for the construction of the SoNaR corpus, hereinafter termed 'Corpus'
2. The License Holder hereby also grants to the NTU and the NTU accepts, the right to multiply (a part of) the Language Data, to publish them, to reuse and utilise them, including a) the right to grant sublicenses to third parties for the use of the Corpus in research and education b) the right to include and/or integrate the Corpus in other systems such as corpora and corpus query systems and c) the right to grant sublicenses to third parties for the use of the Corpus for application development and the use of these applications for commercial and non-commercial purposes.
3. The NTU shall only grant a right to third parties to use the Corpus for the benefit of product development for commercial purposes under the condition that the Products shall not be recognisable in the new products to be developed by these third parties. The NTU shall not give the right to third parties to publish full texts or fragments of the Language Data.
4. The Language Data will only be partially recognizable, i.e. the user will not be able to see or download the entire text. The visibility of the text will be limited to passages. The

