A term extractor for the World Meteorological Organization - a comparison of different systems

Mejia, Sofia

UNIVERSITÉ DE GENÈVE
FACULTÉ DE TRADUCTION ET D'INTERPRÉTATION

MA Thesis

Sofía Magdalena Mejía

A Term Extractor for the World Meteorological Organization – A Comparison of Different Systems

FTI, Université de Genève
Département de traitement informatique multilingue

MA Thesis Director: Prof. Marianne Starlander
MA Thesis Jury: Prof. Lucía Morado Vázquez

Academic year: 2015-2016; session: January 2016

J'affirme avoir pris connaissance des documents d'information et de prévention du plagiat émis par l'Université de Genève et la Faculté de traduction et d'interprétation (notamment la *Directive en matière de plagiat des étudiant-e-s,* le *Règlement d'études de la Faculté de traduction et d'interprétation* ainsi que l'*Aide-mémoire à l'intention des étudiants préparant un mémoire de Ma en traduction*).

J'atteste que ce travail est le fruit d'un travail personnel et a été rédigé de manière autonome.

Je déclare que toutes les sources d'information utilisées sont citées de manière complète et précise, y compris les sources sur Internet.

Je suis conscient-e que le fait de ne pas citer une source ou de ne pas la citer correctement est constitutif de plagiat et que le plagiat est considéré comme une faute grave au sein de l'Université, passible de sanctions.

Au vu de ce qui précède, je déclare sur l'honneur que le présent travail est original.

Nom et prénom : MEJÍA Sofía Magdalena

Lieu / date / signature : Genève, le 10 décembre 2016

****

Acknowledgements

I would like to thank first and foremost my thesis director, Marianne Starlander, for her infinite patience, her support and encouragement. For being always available and for the happiness she exudes. I am also grateful to Lucía Morado Vázquez for accepting to act as jury of this thesis.

From the University of Geneva I would also like to thank Aurélie Picton and Donatella Pullitano for their generous advice on term extractors and for helping me and my director to shape this study.

I would also like to thank the World Meteorological Organization (WMO) for letting me conduct this research with their documents and their tools. For that I thank Maja Carrieri and Alexander Koretski. Special thanks to Antonella Romeo-Magnat and Sarah Eymann with whom I have worked closely at WMO and whose advice has been most precious.

My deepest gratitude goes to my husband, Victor, for his untiring support, and to my family for their constant encouragement. I also want to include my friends who have filled my student days with joy.

To all, thank you very much.

Table of Contents

Acknowledgements - 3 -
1 Introduction - 5 -
2 Term extraction - 7 -
3 Term extraction tools - 16 -
3.1 Common problems of term extractors - 17 -
3.2 TermoStat - 19 -
3.3 MultiTrans Prism - 20 -
3.4 SynchroTerm - 22 -
4 Language Services at the World Meteorological Organization - 24 -
4.1 The Structure - 24 -
4.2 The Work - 25 -
4.3 The People - 26 -
4.4 The Terminology - 27 -
5 CAT tools evaluation - 29 -
5.1 ISO/IEC 25010 - 31 -
5.2 ISO/IEC 25010 and the EAGLES framework - 32 -
5.3 The EAGLES 7-step recipe - 36 -
6 Our evaluation - 38 -
6.1 Execution of the evaluation - 46 -
6.2 Pre-test arrangements - 47 -
6.3 The tests - 52 -
6.4 Test results - 59 -
7 Conclusion - 62 -
Bibliography - 65 -
Annex A – WMO authorization to use their publications and to publish results - 67 -
Annex B – Reference list for sub-corpora C_{1a} and C_{1b} - 68 -
Annex C – Extraction results by TermoStat of sub-corpora C_{1a} and C_{1b} - 71 -
Annex D – Mathematical calculations for sub-corpora C_{1a} and C_{1b} - 77 -
Annex E – Reference list for sub-corpora C_{3a} and C_{3b} - 78 -
Annex F – Extraction results by MuliTrans of sub-corpora C_{3a} and C_{3b} - 84 -
Annex G – Extraction results by SynchroTerm of sub-corpora C_{3a} and C_{3b} - 89 -
Annex H – Sub-corpora C_{1a} and C_{1b} - 94 -

1 Introduction

I have been working for the World Meteorological Organization (WMO) on a temporary basis since June 2014. This organization is a specialized agency of the United Nations and it is "the UN system's authoritative voice on the state and behaviour of the Earth's atmosphere, its interaction with the oceans, the climate it produces and the resulting distribution of water resources" [1], as is stated on its website. As such, this organization deals with a lot of technical texts and produces several technical publications.

The idea to conduct this research has risen from personal interest and from seeing that almost all the input to WMO's terminological database[2] is made on a case-by-case basis, upon the request of translators whenever they came across a term. Manual term extraction is done on some documents. At WMO we would like to extract terminology from larger publications, without having to read the whole publication, and to then import it to our terminology database. We conduct this research to find an automatic extractor that is suitable for WMO context.

Outline

Through the different chapters we will explain what term extraction consists of, the different tools we will compare, the method chosen to analyse these tools and the results.

In Chapter 2 we explain what term extraction is. We give a context for most term extractions and we introduce the notion of terminographers, who are the professionals in charge of performing terminological extractions. We explain the procedure established in terminology discipline in order to arrive to good results and adapt it to the context of this study. Each step of this procedure is explained in detail.

We then move on to describe term extraction tools with a brief history and an explanation on how they work in Chapter 3. We introduce the software that we have used during our research.

[1] https://www.wmo.int/pages/about/index_en.html
[2] METEOTERM: http://wmo.multicorpora.net/MultiTransWeb/Web.mvc

In Chapter 4 we describe the Language Services at the World Meteorological Organization, which is the context for this research. We explain how terminology is managed at WMO.

Chapter 5 describes the history of CAT tools evaluation and then goes on to describe the ISO standards that exist to evaluate them and the framework that was used during our research, the EAGLES report.

In Chapter 6 we describe how we performed the evaluation according to what was established in Chapter 5, all the necessary arrangements to carry out the evaluation and the results.

We conclude our work in Chapter 7, where we analyse the results of our tests and present our recommendation for WMO. We also provide some ideas for further research that could be carried out as a continuation of this MA thesis.

2 Term extraction

Terminography is the practice of terminology. According to (Cabré, 1999), Terminography involves gathering, systematizing and presenting terms from a specific branch of knowledge or human activity. It is an ongoing process which can be separated into several stages, as is shown in Figure 1. Term extraction is a part of the process of the terminographer's work.

The terminographer's work is a terminological search, which can be monolingual, multilingual or monolingual with equivalents, and systematic (when it covers the terms of an entire special subject field or a subpart thereof) or *ad-hoc* (when the research is restricted to a single term or set of terms) (Cabré, 1999).

(Cabré, 1999) proposes a methodology to be followed by terminographers in six stages for the work to be systematic. The first step in her methodology is to define and delimit the task that the terminographer is set out to accomplish. It is very important to establish the scope of the work from the beginning, since not doing so may result in having to re-orient the work halfway through o redefining the entire work which is costly. The main important issues to be defined before the work begins are the topic, which will determine the scope of the work being done; the users of the product, which will determine the documentation needed; the purpose of the work, be it standardizing terminology or describing the terminology of a domain; and, finally the resources that will be needed for completing the task (number of people, financial resources, among others).

The second step in her methodology is preparing for the task. The terminographers have to define the corpus that will be used for the terminographic work, the experts that are going to be consulted, the structure of the field for which they will provide a terminology, and a schedule.

The third step consists of preparing the terms and the information about them that is needed according to the specific task. Within this step terminographers proceed with the term extraction, the extraction record (where they collect the terms as they occurred in the corpus with any other pertinent information), and the terminological record (which structure has to be defined according to the specific task) where they orderly present each term with the information assigned to it.

In the fourth step terminographers decide how they are going to present their work, according to the needs of the users established in the first step.

The fifth step is when the entire task is revised, from the choice of corpus to the information provided in the terminological record and it is usually in this step where neologisms are assessed for their viability.

The last step she proposes is for dealing with unansewered queries. Sometimes the opinion of other experts is needed. The figure below shows a schematic representation of these stages.

Stage 1	**Definition and delimitation of task** topic addressees purposes size
Stage 2	**Preparation of search** acquisition of information choice of consultants choice of information selecting an extraction corpus structuring the field proposing the work schedule
Stage 3	**Preparation of the terminology** extraction extraction record terminological record
Stage 4	**Presentation of work**
Stage 5	**Revision of work**
Stage 6	**Treatment and resolution of problematic cases**

Figure 1: Cabré's stages followed in systematic monolingual and bilingual searches (Cabré, 1999, p. 131).

(L'Homme, 2004) also describes the terminographer's work as starting with a terminological research, which she defines in the same way as (Cabré, 1999); except she uses different names for them: *recherche thématique* for systematic research and *recherche ponctuelle* for *ad-hoc* research.

(L'Homme, 2004) also proposes a set of stages in the terminographer's work, but she distinguishes seven steps instead of the six proposed by (Cabré, 1999). We will present these seven steps in French, with an indicative translation in brackets by us:

1) *La mise en forme d'un corpus* (selecting and building a corpus)
2) *Le repérage des termes* (recognizing the terms)
3) *La collecte de donées* (collecting data concerning the terms)
4) *L'analyse et la synthèse des données* (analysis and review of data collected in the previous stage)
5) *L'encodage des donées* (coding data in the specialized dictionary or terminological database)
6) *L'organisation des données terminologiques* (organizing terminological data by the selected criteria)
7) *La gestion des données terminologiques* (managing terminological data)

The schemes proposed by both authors have a lot in common and the stages are self-explanatory.

We will perform a systematic research and we will base our work on the stages proposed by both authors, creating a scheme of our own, as seen in Figure 2, listing all the stages that are relevant to this project.

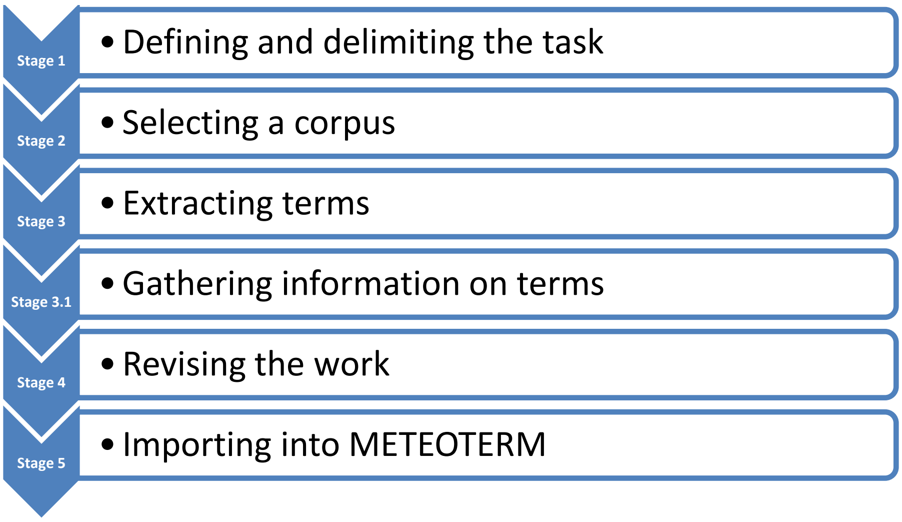

Figure 2: Our stages on the terminographic work related to this study

Stage 1 – Defining and delimiting the task

What we will be doing is enlarging WMO's terminological database, METEOTERM, creating terminological records for a specific domain by performing a monolingual terminological extraction with equivalents (systematic research). It can be seen as part of a bigger terminographic project, where the basics like end-users, purposes, size, design of the terminological record, and conceptual structuring of the field are decisions that have already been taken and assessing them is not within the scope of this research project. This is why there is no need to go into detail for all stages proposed by the above mentioned authors; however, we will go into detail for the stages and sub-stages that are important for our project.

The goal of this research project is to find the most suitable extraction tool for the World Meteorological Organization, so we will be using IT tools as much as we can. According to (L'Homme, 2004), IT has been used for Terminography since the 60s, but at the beginning it was used only for knowledge transfer. Researchers in the field realized that it was impossible to manage terminological data supported in paper, so terminology data banks were born. Even though storage was electronic, terminological researches were done in paper until the mid-80s.

From my personal experience working in different international organizations and attending conferences, a lot of effort is constantly put into making terminological databases more user-friendly, more flexible and more secure. Companies are listening to international organizations and are trying to keep up with the demands, and the industry is evolving very fast, as all IT-related industries.

Stage 2 – Selecting a corpus

According to (Cabré, 1999), the corpus from which terms will be selected must meet a series of conditions in order to ensure the reliability of the results (Cabré, 1999, p. 134):

- it must be pertinent, i.e., representative, of the field being analysed, and, if possible, written by a highly regarded author
- it must be complete, and as such include all aspects of the terminological task to be performed

- it must be up-to-date, so that the list of terms obtained will be useful
- it must be original, i.e., written in the language in which the terminological work is being carried out.

With regards to this last point, this is especially true when standardizing terminology.

As both (L'Homme, 2008) and (Bowker, 2002) agree, the definition of a representative corpus is very difficult to establish, since it depends on the activity for which the corpus will be used. In our particular research context, where the main end-users of the terminological records will be WMO translators, we need to provide terminology that has been approved by WMO. In order to do so, the corpus will be a parallel corpus composed of original texts in English and their translations into Spanish that have been revised and edited. These publications are in electronic format, so they can be processed by a term extractor.

For this project, the corpus will be composed of parts of two WMO publications:

A) *Abridged Final Report with Resolutions and Recommendations of the Extraordinary Session (2014) of the Commission for Basic Systems* (WMO-No. 1140)

The Commission for Basic Systems (CBS) is one of the eight technical commissions. The main activities of the CBS are related to the development, implementation and operation of integrated systems for observing, data processing, data communication and data management, and to the provision of public weather services, in response to requirements of all WMO Programmes and opportunities provided by technological developments. The Commission meets every two years in an ordinary or extraordinary session and after each session a report is compiled with all documents that were discussed during the session and the decisions made. The only parts of this publication that are revised and edited, therefore reliable, in both languages are the Resolutions and Recommendations adopted by the CBS, without their respective annexes. These account for a total of 8,834 words in English and 11,253 words in Spanish. Usually, there is a manual terminological extraction of these parts done once the final version of the report is ready, as is the case for the report of all sessions at WMO.

B) *Guide to Meteorological Instruments and Methods of Observation* (WMO-No. 8)

This is a very large publication and it is one of the most technical publications produced by WMO. Publications are usually outsourced to external senior translators. The 2010 update of this publication was outsourced for translation and revision to the Spanish State Meteorological Agency (AEMET), and a new update has been received and will be outsourced for translation during the next few months. WMO has a double interest in extracting terminology from this publication: on the one hand, to get and acknowledge the terminology used by AEMET; on the other hand, to have it ready for use in METEOTERM for the upcoming translation. A specific chapter of this publication was chosen for this project, that is Chapter 11 of part II (called *Urban Observations*), because according to AEMET it was one of the most difficult chapters to translate due to the terminology. They called in specialists on urban methods of observations and long discussions took place between them and the Real Academia Española[3] and Fundéu[4] (authoritative voices on Spanish). This chapter accounts for 15, 217 words in English and 17, 629 words in Spanish (excluding bibliography).

These two publications share the same domain, that is why we can group them in one corpus, and present two different types of texts, which is an interesting factor when evaluating software, to assess how the software reacts.

Even though it may seem like quite a small a corpus, we judge it appropriate for the specific work that is being done here: evaluating a term extractor for its use at WMO.

Stage 3 – Extracting terms

According to (Cabré, 1999), "terminological extraction (sometimes called excerption) consists of selecting those elements out of the corpus that are considered terms of the special subject field which is the object of the terminology" (p. 136).

A terminological extraction can be done manually or using electronic tools that have been built with this purpose. This seems obvious, but it is worth mentioning it since we will be doing both for this project. We will discuss term extraction tools in Chapter 3. Broadly, the way they work is that we submit a

[3] http://www.rae.es/
[4] http://www.fundeu.es/

corpus into the tool (which can be monolingual or bilingual) and the tool performs the extraction and offers a list of "candidate terms", with their translation in the case of a bilingual corpus. This list must then be assessed by the terminographer.

The most complex part of the term extraction process is deciding whether a term is actually a term and should be included in the terminological database and, in some cases, realizing where the term starts and where it finishes. We may come across a term that belongs to the general language or a term that belongs to a different topic (though it is not likely to happen in our extraction, since the corpus is very restricted).

In order to decide whether we should include a term in our database, we followed the indicators of termhood, cited from p.137 (Cabré, 1999):

- a phrase is lexically organized around a single base (random access memory, central processing unit, communication adaptor unit)
- other linguistic elements cannot be inserted into the terminological phrase (*head of household but not head of the household)
- none of the parts of the phrase can be modified individually (*power of attorney but not power of many attorneys)
- the term can be replaced by a synonym (…)
- an antonym exists in the same special subject field (…)
- the frequency with which the same terminological phrase occurs in texts of a particular special field
- the phrase is a single lexemic unit in other languages (Spanish "tipo de letras" = font, Spanish "portaaviones" = aircraft carrier)
- the meaning of the expression as a whole cannot be deduced from the separate meanings of its parts (…)
- the presence of certain linguistic units inside a phrase indicates that the phrase is most likely a freely combined string (The performance of this jazz *or* soul singer).

(Kageura & Umino, 1996) define termhood as the degree a linguistic unit is related to or represents a domain-specific concept. This concept applies both to simple and complex terms. They also use the concept of "unithood" as the degree of strength or stability of syntagmatic combinations or collocations. This concept applies to simple and complex terms as well as other complex units such as collocations.

We will follow these indicators in our extraction. It will be interesting to see how the different systems recognize and delimit the terms and to evaluate if it saves time. As for deciding which terms are ultimately added to the database, at WMO it is the Language Support Unit officers together with the Spanish revisers that will validate them.

Stage 3.1 – Gathering information on terms

(Cabré, 1999) then explains that for each occurrence of a term an extraction record should be filled in with data concerning the context of each occurrence. Then, a terminology record will be created for each term with the relevant information contained in all extraction records of said term.

She establishes that the selected terms are included in an extraction record with the following details (p. 138 (Cabré, 1999)):

- the canonical form of the term or phrase
- the context, i.e. the part of the text in which the terminological unit functions grammatically
- the reference of the document in which the term is found: its formal and typological characteristics and its contents
- the grammatical category and subcategories (gender, conjugation class, etc.) that are evident from the form the term has in the context in which it appears
- other, more varied and irregular information, such as an equivalent form, an illustration, a complementary definition, etc.
- the information on an extraction record is usually complemented with information about the management of the extraction process: author, data, position
- miscellaneous data

(L'Homme, 2004) suggests using a concordance tool in order to see the terms in context in every occurrence and to be able to gather the information contained in the context.

Term extractors usually show the term in context and our ideal scenario would be an extraction output that we can then import to our database without having to make changes or where the changes are minimal and not very time-consuming. The terminological record has already been established and evaluating it is not within the scope of this research. There is no written guideline as to which fields are compulsory in a WMO terminological record, but the data usually asked from translators when they submit a term is the term in the original language with its definition (if available) and sources and the term in the target language, with its definition (if available), sources and an indication of whether it has been discussed by all in-house translators of that language or if it is a suggestion of only one translator.

The only case in which we will use an extraction record is to perform manual extractions. The use of tools enables terminographers to skip the creation of an extraction record and create directly a terminological record, since all the information of all occurrences of a term are visible at the same time.

Stage 4 – Revising the work

So far, the process described will be carried out by Language Support Unit officers at WMO (structure of the language service will be explained in Chapter 4). In this stage, the work is revised and problematic cases submitted to Spanish revisers for their solution (experts will also be contacted if necessary).

Stage 5 – Importing into METEOTERM

The final product for end-users is WMO's terminological database, METEOTERM, so the last stage is to import the terms that have been extracted.

These are all the stages of our terminographic work. We will describe the work that falls within the scope of this research, which is up to stage 3. The remaining two stages are not relevant for the evaluation of term extraction tools itself, but they are the completion of the terminographic work. The information obtained from this evaluation will be provided to WMO to complete these two last stages, but the processes will not be explained here. The goal is to evaluate the tools that will help us to do this, so we will explain what these tools are and how they work on the next chapter.

3 Term extraction tools

As we mentioned in the previous chapter, IT has been used for Terminography since the 60s, but it wasn't until the 80s that various disciplines expressed a need for extracting terminological units from texts. The first broadly known term detector appeared in 1990 and it was called TERMINO. It was developed at Centre ATO (Analyse de Texte par Ordinateur), a computational linguistics research centre associated with the Université du Québec à Montréal (UQAM) (Cabré, Estopà Bagot, & Vivaldi Palatresi, 2001[5]; Lauriston, 1994).

The goal of term extraction tools is to find words or word combinations that are susceptible of being terms in a given corpus. This is a task that can sometimes be difficult even for a terminographer; that is why we say that term extraction tools produce a list of **candidate terms** that has to be revised by a terminographer in order to decide if these candidates are actually terms or not.

Term extraction tools use several indicators to provide the list of candidate terms (L'Homme, 2004):

- Frequency: usually, a word that is important to a field will be repeated more than words that are not.
- Predominance of terms of nominal nature: a lot of term extractors search only for nouns.
- Complexity of terms: a lot of terms are complex terms (i.e., composed of more than one word) and certain specialists believe that this is the most common term structure. We can also find term extractors that only search for complex terms.
- Finite number of sequences that can constitute a complex term: studies have shown that complex terms are made of a finite number of sequences of parts of speech, so there are also term extractors built to find these sequences.

Term extraction tools use either linguistic, statistical or hybrid knowledge to perform their task (Bowker & Pearson, 2002; Bowker, 2002; Cabré et al., 2001; L'Homme, 2004; TermCoord European Parliament, n.d.).

[5] In (Bourigault, Jacquemin, & L'Homme, 2001) *Recent advances in Computational Terminology*, pages 53-88

Tools that use a linguistic approach try to identify word combinations that match part of speech patterns (in English, for example, many terms consist of NOUN+NOUN or ADJECTIVE+NOUN). The text from which terms are extracted is first tagged, and then the tool recognizes the patterns and offers a list of candidate terms that follow this pattern. The main inconvenient of this type of tools is that they are language specific since term formation patterns vary from one language to another (Bowker, 2002; L'Homme, 2004).

Tools that use a statistical approach identify repeated lexical units within a text. The frequency threshold can usually be specified by the user and there exist stop lists, which are lists of words that the tool will ignore and will not include in the list of term candidates. Almost all term extractors come with a default stop list that can be tailored to users need. The main advantage of this type of tools is that, unlike the ones with a linguistic approach, they are not language specific (Bowker, 2002; L'Homme, 2004).

A related approach to statistical term extraction is to calculate Mutual Information. This means that if two lexical units appear together more often than they appear separately, it is likely that we are in front of a complex term (Bowker, 2002).

Hybrid tools combine both approaches. They mainly use the statistical approach and then apply syntactic rules and filters to select the candidate terms with certain syntactic characteristics; or they generate a list based on linguistic data and then refine the list with statistic calculations (Cabré et al., 2001; L'Homme, 2001, 2004; TermCoord European Parliament, n.d.).

3.1 Common problems of term extractors

The list of candidate terms that term extractors produce is usually not perfect. In this list we may find candidate terms that do not interest the user, and this is called **noise**. It can also happen that the tool does not include in the list some terms that do interest the user, this is called **silence** (Bowker & Pearson, 2002; L'Homme, 2001, 2004).

Noise and silence of a term extractor are measured by **precision** and **recall**. **Precision** is the proportion of terms that are actually terms in the candidate terms list that the tool provides. Precision is high when there is little noise. **Recall** is the proportion of terms extracted of all the terms appearing in the corpus. Recall is high when there is little silence (L'Homme, 2001, 2004).

These measures can be interpreted as the capacity of the tool to extract all terms from the corpus (recall) and the capacity of the tool to differentiate which units are terms and which are not (precision) (Cabré et al., 2001).

(L'Homme, 2004) states that from the point of view of the users, the best option is to use a term extractor that reduces silence to the maximum extent, even if this increases noise significantly: it is easier to clean a list with unwanted candidate terms than to find a term in a large corpus. We agree with this statement, especially for very large corpora.

(L'Homme, 2004) proposes three methods to assess the performance of a term extractor:

- The list of candidate terms is compared to a terminology data bank or a specialized dictionary. This technique allows measuring noise and silence but presupposes that the dictionary or bank against which we are comparing our list contains all the terms of the field we are dealing with, which is almost impossible. A term in the candidate list that does not appear in the dictionary or bank will be considered as noise.
- A terminographer (or any other user) validates the list of candidate terms. This technique measures noise but not silence.
- A terminographer (or any other user) performs an extraction and provides a reference list. This list is then compared to the list of candidate terms. In this case it is difficult to know if the reference list contains all the terms of a specialised text or only part of them

We will be using the third technique to measure precision and recall in order to arrive to our conclusion.

Nowadays there is a great variety of term extractors using different approaches, some of them can only be used for research purposes, some of them are a module of a larger linguistic solution and some of them are online and free.

Currently there is a generalized belief among terminologists and chiefs of linguistic services in international organizations that no term extractor available in the market is good and that using one will result in a waste of time. From our experience, most of them have reached this conclusion without having done the necessary tests. In the cases where some testing took place, only a few followed a methodology and reached a valid conclusion. In most cases tests were done without the necessary planning or the necessary time. Finding the resources to test new technology in international organizations is a major problem in the area, and that is why it takes so long for them to update to current trends and advances.

We have found an international organization interested in testing software, and for that we consider ourselves very lucky.

In the next paragraphs we will introduce the software we will use for our research.

3.2 TermoStat[6]

TermoStat was developed by Patrick Drouin at the University of Montreal. It is a hybrid automatic term extraction tool which method is to compare a domain-specific, technical corpus to a non-technical corpus in order to identify terms. Terms that appear "abnormally" more frequently in the technical corpus than in the non-technical corpus are more likely to be terms belonging to the specific domain of the technical corpus (Drouin, 2003).

After analysing the submitted corpus, TermoStat provides a list of candidate terms with a score based on the frequency of the term in the analysed corpus and on the frequency of the term in the reference corpus.

TermoStat performs monolingual extractions in French, English, Spanish, Italian and Portuguese.

The English reference corpus has 8 000 000 occurrences which correspond to approximately 465 000 different forms. Half of this non-technical corpus is composed of newspaper articles with varied subjects from the daily newspaper *The Gazette* of Montreal, published between March and May 1989. The other half of the corpus comes from the British National Corpus.

The Spanish reference corpus has about 30 000 000 occurrences which correspond to 527 000 different forms. It is composed of texts from the European Parliamentary Assembly.

[6] Information obtained from the help section at: http://termostat.ling.umontreal.ca/index.php

TermoStat works in three stages:

1. Tagging

The texts submitted to the tool are tagged by an external tagging tool called TreeTagger. The goal of this stage is to disambiguate the words that could have more than one syntactic characteristic. Every word in the corpus will have a syntactic tag by the end of this stage.

2. Extracting character strings that match a predefined set of rules

TermoStat applies a filter to the tagged text that contains regular expressions to extract words or set of words that match the predefined syntactic matrices.

3. Weighting and selecting term candidates

Every term candidate receives a score according to the method chosen when displaying the results. The candidates with the higher scores are considered the most relevant in the text. An acceptability threshold allows the user to exclude certain words or expressions not considered terminological in the text.

With advice of experts at the University of Geneva, we have chosen this software to help us build the reference list that will be used for our research. In Chapter 6 we will explain how we will proceed.

3.3 MultiTrans Prism[7]

MultiTrans Prism is a Modular Translation Management System recently acquired by R. R. Donnelley. It is based on two components, a TextBase, which is a corpus-based translation memory (i.e. it provides full-text searches, context is always available), and a TermBase, which is the terminology management component.

The TextBase is where users stock the texts that have already been translated with their translations to use as reference when translating new texts. Translations are aligned automatically by the tool and these alignments can be modified by the user. The tool provides a Translation Agent which automatically

[7] http://rrdonnelley.com/languagesolutions/solutions/multitrans.aspx

shows the user the segments that match the segment being translated. The TextBase allows seeing the segment in context.

The TermBase module allows the users to create one or several terminological databases, according to their needs. One function of this module is to extract terminology from TextBases.

There are two ways to extract terminology in MultiTrans 5.5, through the translation generator or through the Analysis Agent[8].

MultiTrans performs both monolingual and bilingual extractions: it extracts recurring simple and complex candidate terms and provides statistics on the frequency of said elements. Then algorithms analyse the extraction and identify and retrieve probable corresponding translations (Gervais, Blvd, & Jy, n.d.).

We have been unable to find further information on whether they use another approach combined with the statistical approach.

This tool actually performs a 'monolingual extraction with equivalents' (following Cabré's terms), since it first extracts the terms from source language and then it produces a short list of possible translations for each term. It does not analyse the target corpus for extraction, it is based on the alignment the tool makes of source and target texts.

To perform an extraction in MultiTrans, the first step is to upload the corpus from which we want to extract the terminology (we build a TextBase). Then we indicate the tool to perform the extraction, and it produces a list of candidate terms in the source language. Then we indicate the tool to generate the translations of those terms and the tool provides possible translations for those candidate terms together with the context for each occurrence of the candidate term. The tool creates a 'working termbase' where we can insert the terms we validate with their translations. We can import one term at a time or many terms at the same time. When a candidate term is imported to this working termbase, the symbol of a paper clip appears next to it, to indicate that the term already exists in the termbase.

In Figure 3 we show a screenshot of the working environment of this tool.

[8] Unfortunately, we were not able to test the MultiTrans' Analysis Agent because of technical issues. These issues were signalled to R.R. Donnelley in July 2015 and by the time this work was published the issues had not been solved.

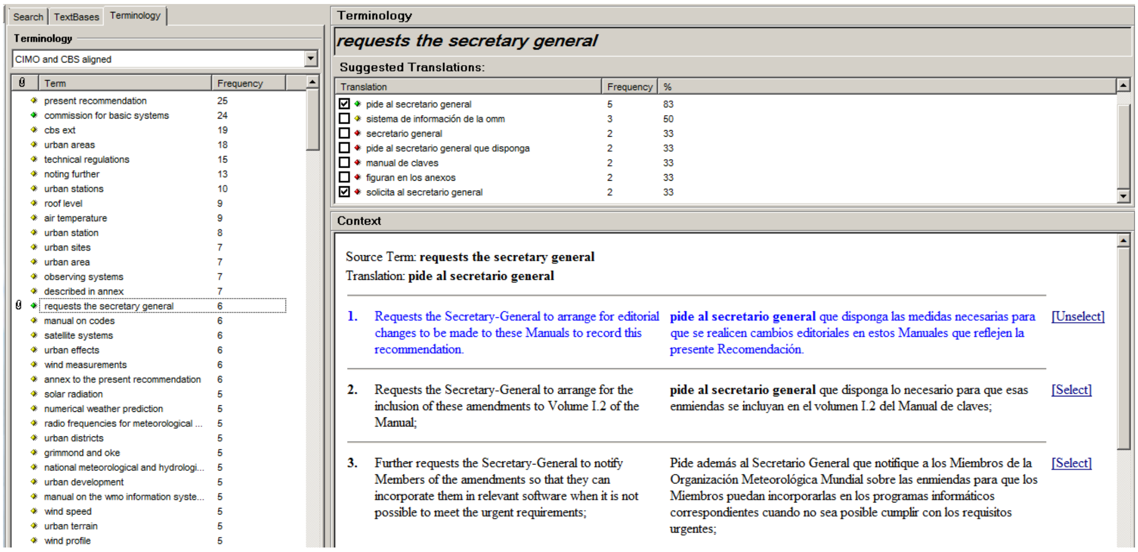

Figure 3: MultiTrans translation generator – working environment.

We are testing this software because it is the current software at WMO.

3.4 SynchroTerm[9]

SynchroTerm is a product of the Canadian-based company Terminotix. The latest version was launched in 2014.

SynchroTerm is a bilingual term extractor and works in two stages. First, it scans two halves of the input files to compile a list of frequently occurring source and target language expressions. Then, it applies statistical, syntactical and morphological algorithms to find possible equivalents for the source expressions.

To perform an extraction with this tool the first step is to upload the corpus from which we want to extract the terms. Then we indicate the tool to perform the extraction and it offers a list of candidate terms with possible equivalents in the target language with a percentage of probability that that target term is the correct equivalent of the source term. The tool also shows the context of occurrence for each term.

With every extraction project, the tool creates a terminological database, a 'working termbase', were validated terms with their translations can be inserted. We can import one term at a time or many at the same time. Every time a term is imported into the working termbase, it changes colour and it is italicized in the candidate term list, to indicate that the term already exists in the termbase.

[9] http://www.terminotix.com/index.asp?name=SynchroTerm&content=item&brand=4&item=7&lang=en

Figure 4 shows the working environment for this tool.

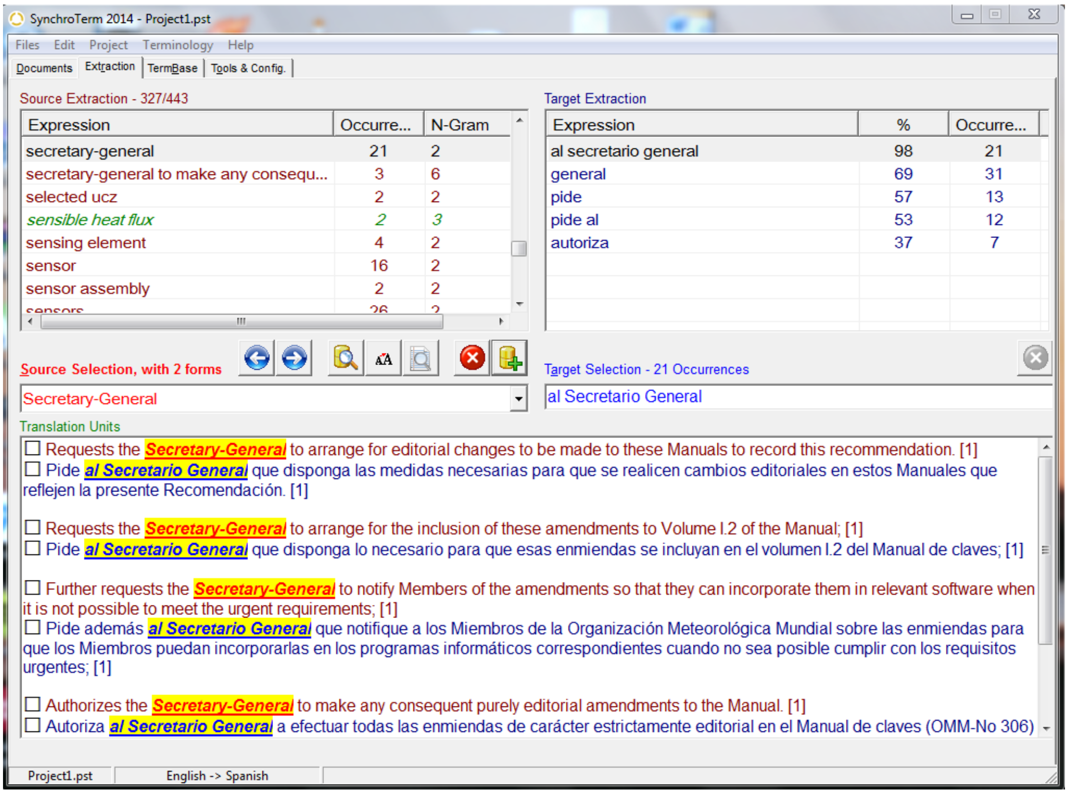

Figure 4: SynchroTerm working environment

We have chosen to test this software because we believe that it has not been widely tested[10]. Our previous experience with Terminotix products has been successful so it is also a personal interest to test Terminotix products.

These are the three tools that will be used for our research. As explained, the first one will be used to build the reference list against which we will compare the extraction results of the second and third tool, which are the actual subjects of the evaluation. Unfortunately, the scope of this research does not allow us to test all the term extractors available in the market.

Now that the tools have been chosen, we can move on to explain the context for which they will be tested in the next Chapter.

[10] It has been tested by the Terminology Coordination Unit of the European Parliament: http://termcoord.eu/useful-links/free-term-extractors/term-extraction-analysis-done-termcoord/

4 Language Services at the World Meteorological Organization

As stated in the introduction, this research takes place in the context of the World Meteorological Organization (WMO). In this Chapter we will explain the structure of the department that deals with translations and terminology management and we will establish the context for our software evaluation.

4.1 The Structure

At WMO there is a Language, Conference and Publishing Services Department (LCP), headed by a Director and a Documentation and Publications Management Chief. This department is subdivided into Client Relationship and Delivery Management (with shared supervision with the Information Technology and Common Services Division), Conference Services with their respective Chief, Documentation and Publications Management Unit, and Language Services and Publications Production.

Conference Services is subdivided into Reproduction Services and WMO Publications Sales and Distribution.

The Documentation and Publications Management Unit, is composed of a group of publishing assistants (the pool and desktop publishers for Arabic, English, French, Russian and Spanish –for Chinese these tasks are outsourced–), a section of Documentation and Publications Preparation (in charge of overseeing that each document goes through the entire workflow), a Design Officer (in charge of creating the design for all WMO publications events and gift shop, among other things), and an Outsourcing Unit (in charge of outsourcing approximately two-thirds of the translation volume).

Language Services and Publications Production is composed of all translators and editors for the six UN languages and a Linguistic Services Support Unit (who is in charge of distributing translations in-house, CAT tools, terminology, references and providing linguistic support as needed).

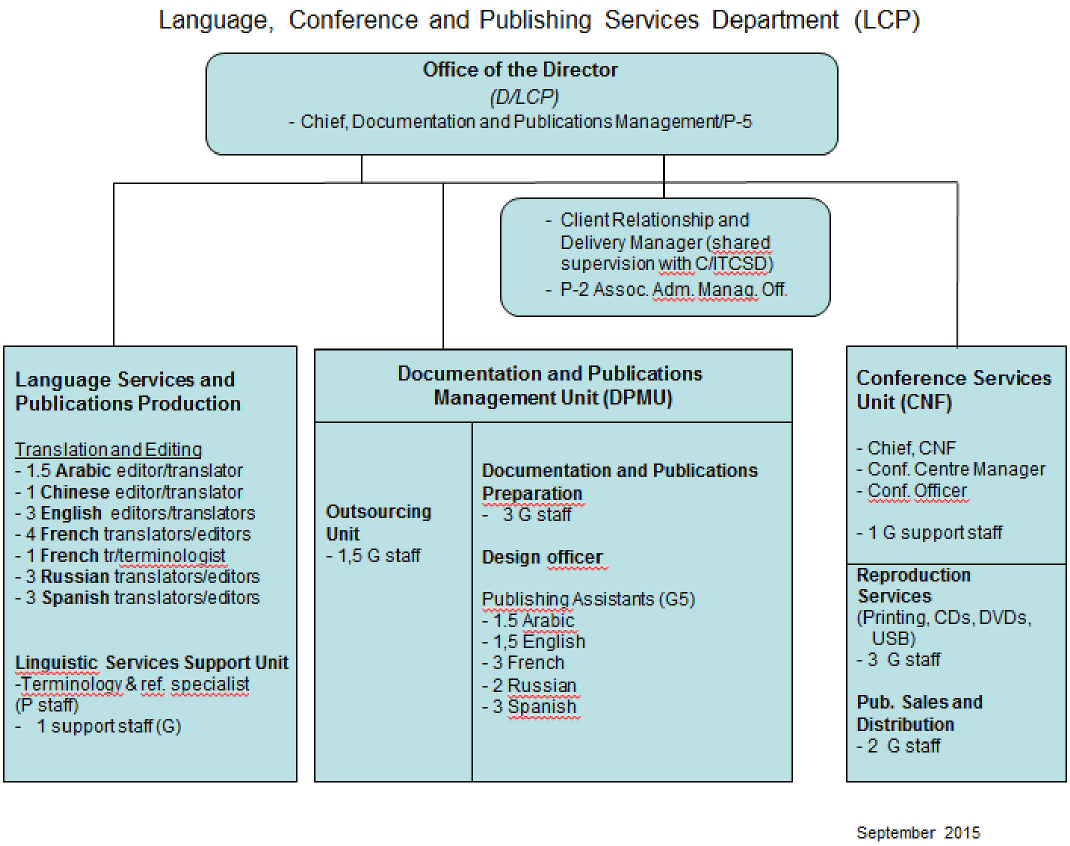

Figure 5: Organization Chart for Language Services at WMO[11]

4.2 The Work

We are going to focus on the Units and teams with which we are doing this research, and those are the Spanish translators and the Linguistic Services Support Unit (LSU).

There are different types of texts translated at WMO:

- There is a large volume of correspondence translated daily, where vocabulary can sometimes be technical.
- Job vacancies and other documents from the Human Resources Department are translated in-house into French.
- Session documents, which are translated before and during the sessions. In these documents, for resolutions and descriptions of activities, we find the

[11] Staff at International Organizations is divided into G staff, where G stands for general services, and P staff, where P stands for professional services. Professional posts are those which require a university degree or its equivalent in a combination of education, training and experience, and which are subject to international recruitment. General Service posts are those which do not normally require a university degree or its equivalent and which are normally filled through local recruitment. Source: WMO Staff Rule 121.1 – Classification of posts and staff.

language that is so characteristic of International Organizations and we can find very technical vocabulary sprinkled in this "onusian" jargon because the bodies of WMO deal with very technical subjects. After each session, all these documents become the session report.

- Publications, which are usually very large and very technical. In general, these documents are outsourced to senior translators.
- Other miscellaneous documents (press releases, web content, etc.).

LSU is primarily in charge of distributing translations done in-house. All translations are pre-treated to check for existing translations, and special references are gathered if necessary. They also provide support to translators and revisers whether they work in-house or from home and assistance with tools when needed. They manage the linguistic resources, the bilingual corpora used for reference, other references, and the terminology database. They also welcome temporary translators and provide them with a crash course on tools and workflow at WMO.

Unfortunately, little support is requested from the Conference Services Unit.

4.3 The People

LCP Department is composed of 46 staff members[12]. Temporary staff members are hired on peak periods, especially during the Organization's sessions.

The Spanish team is composed of three translators/editors. Since the team is so small, between the three of them they take care of several tasks: translating, revising, editing publications, and managing part of WMO website. They also work very close to LSU for terminology issues.

There are two staff members for Spanish pool, who are in charge of formatting the translations once they are revised and edited.

There is one desktop publisher for Spanish who is in charge of getting the publications in Spanish ready for printing.

LSU is composed of three people: an assistant, who has been working at WMO for 27 years and who helps with the alignment of publications to feed into the TextBases and with the pre-treatment of translations; a junior French

[12] Until 7th November 2015.

translator/terminologist who has been working for 4 years at the Organization; and a senior terminologist who has been working for 25 years at WMO who also deals with the distribution of translation work and the management of CAT tools.

4.4 The Terminology

The tool chosen by WMO in 2009 was MultiTrans. There are several TextBases available to translators and a TermBase called METEOTERM, which is publicly available. Most users access the TextBases and the TermBase through the web-based version of the software. LSU officers have the desktop version installed in their workstations for management purposes.

The previous system used at WMO was MetaRead, and import into METEOTERM was done by MultiTrans staff. As with every import, some errors occurred and we still come across them today, though there is a great effort being made to clean the database.

At the moment of importing METEOTERM into the new tool it was decided that the database would not be structured by domain since all domains and subjects dealt with at WMO are very interrelated.

As mentioned earlier, additions to METEOTERM are mostly done by request of in-house translators. It is usually an LSU officer that receives the request, asks for the research the translator has done and then adds the entry (or modification of an entry) to the termbase. The LSU officer also checks that the entry in question has been discussed by all translators of the same language and then suggests translators of the rest of the languages to contribute.

The other mentioned way in which terms are added into the database is by manual extraction of terms contained within resolutions and recommendations in the session reports. The resolutions and recommendations of these reports are edited after their translation, which is why they are considered reliable in terms of terminology –everything is verified. In this case it is the junior LSU officer who performs the extraction and then consults with translators of all languages for approval and agreement (especially for languages that said LSU officer does not manage, like Chinese, Arabic and Russian). The LSU officer looks for specific terms in these reports, and those are the titles of publications, the names of working groups, task teams, workshops held at WMO, commissions created, panels and institutional terms of the like. Then, terms are inserted into METEOTERM.

Some years ago, the previous tools in place allowed for some terminology problems to be pointed out to translators or solved beforehand. Nowadays, unfortunately, there are no resources to deal with terminology in advance, i.e. before the translation is assigned to a translator.

This research is done in an attempt to find a tool that enables us to recover the terminology from translated publications and have it ready for future updates.

5 CAT tools evaluation

Term extractors also fall into the category of CAT tools (computer-assisted translation tools).

In the literature we have found general agreement with the fact that tool evaluations should be carried out in a determined context, since what is expected of a tool may change if, for example, the end-users change. That is why in the previous Chapter we have explained the structure of the linguistic services at WMO.

Currently there is no standard method for evaluating CAT tools, as (Quah, 2006) agrees, due mainly to the variety of tools and of groups testing those tools, though there have been several efforts made towards building a framework for CAT tools evaluation.

When taking a look at the history of evaluation of natural-language processing tools we find that testing has mostly been done for machine translation and for CAT tools other than term extractors (translator workbenches, spell-checkers, grammar-checkers, for example), especially when trying to establish a framework. Nevertheless, there have been tests done for term extractors: (Benavent & Parrilla, n.d.; Cabré et al., 2001; Drouin, 2003; Estopà Bagot, 1999; Jacquemin, 2001; Kageura & Umino, 1996; King, 1992; L'Homme, 2001; Lauriston, 1994, 1995; Rubio, Pastor, & Valero, 2009; Sauron, 2002).

The first milestone in the evaluation of natural-language processing tools was probably a famous report called the ALPAC report, released in 1966 by the Automatic Language Processing Advisory Committee (ALPAC). It studied machine translation in US agencies, among other things[13]. Later, Van Slype established a methodology of evaluation in 1979 for machine translation systems and then Lehrberger and Bourbeau developed yet another methodology in 1988, also for machine translation systems. JEIDA (Japan Electronic Industry Development Association) or JEITA (Japanese electronics and Information Technology Association) as it has been known since 2000, contributed to machine translation evaluation by developing comprehensive questionnaires for end-users in 1992.

[13] This report is commonly referred to as the infamous ALPAC report. One of the main consequences of this report was that it put an end to funding research on machine translation in the US for about 20 years. For more information please see (Hutchins, 1996)

Another big project was the DARPA (Defense Advanced Research Projects Agency) evaluation of machine translation conducted between 1992 and 1994 (Quah, 2006).

As for CAT tools evaluations, when they are carried out in the development stages, they are usually confidential. The ones that are public are generally reviews for end-users carried out when the tools are already in the market. One report that tried to build a framework for the evaluation of translation memory systems and translator's workbenches is the one published by EAGLES (Expert Advisory Group on Language Engineering Standards) in 1996 which took as starting point an international standard for software quality.

The International Organization for Standardization (ISO) is the world's largest developer of standards which 'give world-class specifications for products, services and systems, to ensure quality, safety and efficiency'[14]. In 1991 it developed two series of standards which provided a framework for the quality and evaluation of software (including natural-language processing tools): ISO 9126 (Software Product Quality) and ISO 14598 (Software Product Evaluation). These standards were revised and have been replaced by ISO/IEC 25000 series known as SQuaRE series (Systems and software Quality Requirements and Evaluation).

What EAGLES did in their report was to adapt ISO standards to the translation environment, creating a flexible and modifiable evaluation framework, in which features and attributes could be combined to reflect the needs of end-users (Quah, 2006). The work started by EAGLES was carried forward by ISLE (International Standards for Language Engineering) and TEMAA (Test-bed Study of Evaluation Methodologies: Authoring Aids). FEMTI (Framework for the Evaluation of Machine Translation), another initiative, was an extension of ISLE (Quah, 2006).

In this research project we will follow the EAGLES methodology because it is the most influential initiative and because unlike the other initiatives mentioned here, it is focused on CAT tools rather than on machine translation. We will adapt this framework to evaluate term extractors.

[14] http://www.iso.org/iso/home/about.htm

5.1 ISO/IEC 25010

In this section we will briefly introduce the new ISO/IEC 25010Standard. Further information will be provided in the next section.

As defined in ISO/IEC 25010 (SQuaRE), 'the quality of a system is the degree to which the system satisfies the stated and implied needs of its various stakeholders, and thus provides value'(ISO/IEC 25010, 2011, p. 2). These needs are represented in this Standard by quality models that characterize product quality into characteristics and subcharacteristics.

Currently there are three quality models defined in the SQuaRE series: the quality in use model and the product quality model in ISO/IES 25010 and the data quality model in ISO/IEC 25012. The full set of quality characteristics will not be relevant to all stakeholders, so we will define them and then select those which are relevant to this research.

Quality in use model

The quality in use model characterizes the impact that a system or software product has on stakeholders. It is 'the degree to which a product or system can be used by specific users to meet their needs to achieve specific goals with effectiveness, efficiency, freedom from risk and satisfaction in specific contexts of use' (ISO/IEC 25010, 2011, p. 8). It is determined by factors related to the interaction of stakeholders with the system, like software, hardware, characteristics of the users, tasks and social environment. The five characteristics it defines are: effectiveness, efficiency, satisfaction, freedom from risk and context coverage. We will focus on the intrinsic quality of the software we are testing, that is why we will not be using the quality in use model in our methodology.

Product quality model

The product quality model categorizes system/software product quality properties into eight characteristics: functional suitability, performance efficiency, compatibility, usability, reliability, security, maintainability and portability. Each characteristic has related subcharacteristics which we will explain in section 5.2. This model can be applied to just a software product or a computer system that includes software. In our research it will be applied to a software product.

Data quality model

Data quality is one of the most important components of the quality and usefulness of information derived from that data. The model can be used to establish data quality requirements, defined data quality measures or plan and perform data quality evaluations. Data in this Standards refers to all type of data (character strings, dates, numbers, images, etc), and relationships between data, but it does not include metadata and does not provide guidelines on physical organization of data. This model categorizes quality attributes into fifteen characteristics. This model falls out of the scope of our evaluation.

In our research we are interested by the product quality model since we will be assessing the intrinsic characteristic of two pieces of software.

5.2 ISO/IEC 25010 and the EAGLES framework

In this section we will explain the EAGLES report output, which we will apply to our tests.

Above all, let us keep in mind that this framework was developed with ISO/IEC 9126 as a starting point and what we will do during our research is to follow the methodology proposed therein while taking into consideration the modifications made by ISO/IEC 25010.

Standard ISO/IEC 9126:1991 was concerned with the definition of quality characteristics to be used in the evaluation of software products. These characteristics are the top-level of a hierarchical organization of attributes, which means that they can be broken down into quality subcharacteristics which can be further broken down. ISO/IEC 9126 proposed six quality characteristics and the new SQuaRE series proposes eight. In the next page we copy a table from standard ISO/IEC 25010 that compares the quality characteristics and subcharacteristics of both standards:

ISO/IEC 25010	ISO/IEC 9126-1	Notes
Product quality	Internal and external quality	Internal and external quality combined as product quality
Functional suitability	**Functionality**	New name is more accurate, and avoids confusion with other meanings of 'functionality'
Functional completeness		Coverage of the stated needs
Functional correctness	Accuracy	More general than accuracy
Functional appropriateness	Suitability	Coverage of the implied needs
	Interoperability	Moved to Compatibility
	Security	Now a characteristic
Performance efficiency	**Efficiency**	Renamed to avoid conflicting with the definition of efficiency in ISO/IEC 25062
Time behaviour	Time behaviour	
Resource utilization	Resource utilization	
Capacity		New subcharacteristic (particularly relevant to computer systems)
Compatibility		New characteristic
Co-existance	Co-existance	Moved from Portability
Interoperability		Moved from Functionality
Usability		Implicit quality issue made explicit
Appropriateness recognizability	Understandability	New name is more accurate
Learnability	Learnability	
Operability	Operability	
User error protection		New subcharacteristic (particularly important to achieve freedom from risk)
User interface aesthetics	Attractiveness	New name is more accurate
Accessibility		New subcharacteristic
Reliability	**Reliability**	
Maturity	Maturity	
Availability		New subcharacteristic
Fault tolerance	Fault tolerance	
Recoverability	Recoverability	

Security	**Security**	No previous subcharacteristics
Confidentiality		idem
Integrity		idem
Non-repudiation		idem
Accountability		idem
Authenticity		idem
Maintainability	**Maintainability**	
Modularity		New subcharacteristic
Reusability		New subcharacteristic
Analysability	Analysability	
Modifiability	Stability	More accurate name combining changeability and stability
Testability	Testability	
Portability	**Portability**	
Adaptability	Adaptability	
Installability	Installability	
	Co-existance	Moved to Compatibility
Replaceability	Replaceability	

Table 1: Comparison of quality characteristics between ISO/IEC 9126 and ISO/IEC 25010 (ISO/IEC 25010, 2011, pp. 22-23)

The quality characteristics of ISO/IEC 9126 were accompanied by guidelines for their use, which broke down into three stages: quality requirements definition (stated or implied needs and the ISO standard), evaluation preparation (the selection of appropriate metrics, a rating level definition and the definition of assessment criteria), and the evaluation procedure (measurement, rating and assessment).

EAGLES took these quality characteristics with their guidelines and developed a framework from a user-oriented perspective which has three main components: a set of attributes (there must be sufficient attributes to express all requirements listed by an end-user); the requirements (the needs for which the system is designed which can be functional or non-functional); and the method of evaluation (test types, instruments and test materials) (Quah, 2006).

The quality characteristics of ISO standard will help us identify the requirements, which will be eventually expressed in terms of attributes. They are defined as follows in the SQuaRE series (ISO/IEC 25010, 2011, pp. 10-15):

- **Functional suitability:** degree to which a product or system provides functions that meet stated and implied needs when used in specified conditions.
- **Performance efficiency:** performance relative to the amount of resources used under stated conditions.
- **Compatibility:** degree to which a product, system or component can exchange information with other products, systems or components, and/or perform its required functions, while sharing the same hardware or software environment.
- **Usability:** degree to which a product or system can be used by specified users to achieve specified goals with effectiveness, efficiency and satisfaction in a specified context of use.
- **Reliability:** degree to which a system, product or component performs specified functions under specified conditions for a specified period of time.
- **Security:** degree to which a product or system protects information and data so that persons or other products or systems have the degree of data access appropriate to their types and levels of authorization.
- **Maintainability:** degree of effectiveness and efficiency with which a product or system can be modified by the intended maintainers.
- **Portability:** degree of effectiveness and efficiency with which a system, product or component can be transferred from one hardware, software or other operational or usage environment to another.

We will go into further detail of these characteristics and explain the subcharacteristics in Chapter 6.

As for the third component of the EAGLES framework, the method of evaluation, it is composed of three sub-components: test types, instruments and test materials. The three major test types are scenario tests, systematic tests and feature inspection.

In scenario tests, the main objective is to assess the suitability of a software product for everyday routines. They can be carried out in the normal working place of the user (field test) or in a laboratory (laboratory test).

In systematic testing, the main goal is to test the behaviour of the software under specific conditions with expected results. It can be performed by user representatives and it seeks to avoid human interaction (which is present in scenario tests). It can be task-oriented (to examine whether a piece of software actually fulfils pre-defined tasks), menu-oriented (to test each program feature or function in sequence) or benchmark testing (to examine the performance of systems and check whether a product meets user requirements).

The aim of feature inspection is to be able to describe the technical features of a piece of software in as much detail as possible, so as to allow comparison between systems of the same type (EAGLES, 1996, p. 34). The main instruments in this type of testing are feature checklists. This is primarily the type of test we will be carrying out.

The collection of data can be performed manually (with the aid of questionnaires, checklists, interviews and observations) or automatically (with logging programs and playback programs which record user interaction with the software and allow playback for later analysis). The test materials are usually electronic corpora (although this depends on the reasons for the test being done).

Some years later after the publication of the report, EAGLES produced a short document where they present the 7 major steps necessary to carry out a successful evaluation of language technology systems or components (EAGLES, 1999). This document published in 1999 is called 'The EAGLES 7-step recipe' and we will use it to guide us through our software evaluation.

5.3 The EAGLES 7-step recipe

As the title indicates, this document presents the process necessary to carry out an evaluation in 7 seven steps, mainly by posing questions to the evaluators.

The first step of this recipe is to establish the reasons for which this evaluation is being done, the purpose of the evaluation and to define what is being evaluated.

The second step is to elaborate a task model in order to define all users and the context of use of the software.

The third step is to define the top-level characteristics, assess which features of the software are going to be evaluated and if they are all equally important.

The fourth step consists of producing detailed requirements for the software being evaluated in the light of previous steps, i.e. taking into account the users of the software, context and top-level characteristics.

The fifth step consists of devising the metrics that will be applied to measure the features and requirements established in the previous step.

The sixth step consists of designing the execution of the evaluation and the seventh consists on actually carrying out the evaluation and making the measurements.

In this Chapter we have reviewed the history of CAT tools evaluation and thanks to the SQuaRE series standard and the EAGLES report output we have been able to define the framework for our evaluation.

In the next chapter we will explain in detail our evaluation following the EAGLES framework and 7-step recipe.

6 Our evaluation

In this chapter we are going to explain the evaluation we will be carrying out following the EAGLES 7-step recipe.

We will be performing a black-box evaluation, where we will test the overall performance of the systems without looking at the internal aspects or specific components of the systems (an evaluation of the internal aspects requires knowledge of programming codes and algorithms and is usually known as glass-box evaluation) (Quah, 2006).

This evaluation is being done because at WMO we want to extract terminology from large publications as automatically as possible. We want to have the terminology of certain publications ready for when the next editions of those publications arrive and have to be sent to translation. This way, translators will not have to research again terminology which has already been approved (and, in the case the translator changes from one edition to another, this will help us ensure terminological coherence). The purpose is to find the best term extractor for WMO. Unfortunately, the scope of this research does not allow us to evaluate every term extractor available, so we have chosen, as explained in Chapter 3, MultiTrans Prism and SynchroTerm. We are evaluating them, for the time being, for their extraction in English and Spanish, but we have to keep in mind that this Organization works with UN's six official languages.

MultiTrans was acquired by WMO in 2009. As previously explained, this software is a linguistic solution that has other functionalities as well as term extraction (see section 3.3). There are two ways to extract terminology with this software, through the translation generator or through the analysis agent, though we will only be testing the translator generator[15].

SynchroTerm is a term extractor and is a piece of software that belongs to the linguistic solution ensemble provided by Terminotix.

As for the task model that we have to elaborate, the direct users are LSU officers (from the Linguistic Support Unit[16]). Translators are indirect users, since

[15] As explained in section 3.3, the Analysis Agent at WMO presented technical issues which made it impossible to test. These issues were signalled to R.R. Donnelley in July 2015 and had not been solved by the time this research was published.

[16] See structure of linguistic services at WMO in Chapter 4.

they will be working with the final output of the software, but will have no direct contact with the tool.

The two LSU officers will be performing the extractions in the future. These users can qualify as expert users and capable of using almost any tool with ease (the same goes for the person performing the tests, a temporary LSU officer), that is why it is not a requirement for them that the tool be easy to use, hence we will not be evaluating the usability of these tools.

The test materials are going to be the publications described in Chapter 2.

Next we have to define what we will be evaluating and how to measure those features. These features will be based on the requirements of LSU officers. Our starting point is the quality characteristics and subcharacteristics of standard ISO/IEC 25010.

According to the requirements of LSU officers, the above mentioned tools will be tested for the following:

- Functional suitability
- Compatibility
- Maintainability
- Performance efficiency

These top-level characteristics will be broken down into subcharacteristics and then into measurable attributes.

Functional suitability

With this characteristic we ask ourselves if the functions of the software meet the needs of LSU officers. This characteristic presents two subcharacteristics: **functional completeness** and **functional correctness**. They relate to whether the functions cover the specified tasks and user objectives and to whether the system provides the correct results with the needed degree of precision, respectively.

The main requirements of LSU officers is that the tool is able to perform a bilingual term extraction, that it supports all six UN official languages and that it supports different file formats, mainly word documents and pdf files since most publications are available in those formats. It is also very important to be able to get the source of each term in the extraction process, since we will need it to complete the terminological record. In order to be able to verify this, we have thought of using both publications at the same time when we perform the extraction, to check if the tools recognize the different sources. Another interesting aspect is to see how the tools work when there are images or tables in the corpus, whether they block the extraction process or not and whether terminology is extracted from them.

For most of the attributes we will be measuring here, we will perform a feature inspection test by answering yes/no questions. For functional correctness we will be measuring terminological precision and recall as defined in section 3.1 and we will perform a task-oriented test.

Feature	MultiTrans Translation Generator	SynchroTerm
Can we do a bilingual extraction with the software?	yes/no	yes/no
Does the software support pdf files?	yes/no	yes/no
Does the software support word files (doc and docx)?	yes/no	yes/no
Does the software support all six UN languages (Arabic, Chinese, English, French, Russian, and Spanish)?	yes/no	yes/no
Does it support French as interface language?	yes/no	yes/no
Can we see the extracted term in context?	yes/no	yes/no
Does the software provide a source for each extracted term?	yes/no	yes/no
Can the software process tables and images in the corpus?	yes/no	yes/no
Is the software capable of extracting terminology from tables?	yes/no	yes/no

Table 2: Tests for functional suitability

Precision and recall

We believe that these are the most important features of all, because, in the end, they measure the quality of the extraction. We also believe that it is here where we will find the most differences between both tools.

As we said in section 3.1, **precision** is the proportion of terms that are actually terms in the candidate terms list that the tool provides, and it is high when there is little noise. **Recall** is the proportion of terms extracted of all the terms appearing in the corpus, and it is high when there is little silence (L'Homme, 2001, 2004).

	Term	Non term
Extracted	TP (true positive)	FP (false positive)
Not extracted	FN (false negative)	TN (true negative)

Table 3: Contingency table to calculate precision and recall taken from (Bernier-Colborne, 2012)**.**

To calculate these measures we will use the formulas P=TP/(TP+FP), where P stands for precision, and R=TP/(TP+FN), where R stands for recall. Ideally, both values are equal to 1 for a perfect precision and perfect recall.

These measures can be interpreted as the capacity of the tool to extract all terms from the corpus (recall) and the capacity of the tool to differentiate which units are terms and which are not (precision) (Cabré et al., 2001).

It is worth noting that they will be calculated from the extraction results obtained in English, since the tools perform firstly the extraction in English and then search for equivalents in Spanish.

Compatibility

Here we want to test if the tools can exchange information with other software. The extraction results will ultimately have to be imported into METEOTERM. It seems quite obvious that in this test MultiTrans will have the advantage since it is normal that two pieces of software from the same developers can communicate and it is, actually, the case: we can import into METEOTERM directly from the translation generator and from the analysis agent. So what we will be testing is if they can export the extraction results into Excel or into an international exchange format. This way, no tool will have an advantage and it is a useful test should METEOTERM ever change its platform.

We will also be performing a feature inspection test by answering yes/no questions:

Feature	MultiTrans Translation Generator	SynchroTerm
Does the tool export results in Excel?	yes/no	yes/no

Does the tool export results in an international exchange format (tbx, xml, html)?	yes/no	yes/no

Table 4: Tests for compatibility

Maintainability

The goal here is to test how effectively and efficiently we can modify the tools. We will focus on subcharacteristic **modifiability** to check if the modifications LSU officers might need to make introduce defects on the quality of the tools. We want to test two different types of modifications, the ones we make prior to the extraction process and the ones we make over the extraction results. Usually, extraction tools provide an 'ignore list', 'exclude list' or 'stop list', i.e. a list of words that should not be included in the candidate terms list because they are common words, such as prepositions, for example. It is interesting for LSU users to be able to modify this list and add words they deem relevant to exclude. After the extraction, it is interesting for LSU users to be able to modify the extraction results if, for example, a term made up of five words was erroneously extracted as containing four words or if the equivalent in Spanish was erroneously extracted. We also would like to measure if this modifications are easy to make or not, and we will do this by counting the number of clicks needed to introduce these changes.

Once again we will be performing a feature inspection test with the help of the following questions:

Feature	MultiTrans Translation Generator	SynchroTerm
Can we customize the software (does it have an ignore list)?	yes/no	yes/no
Can we edit the ignore list?	yes/no	yes/no
How easy difficult is it to edit the ignore list?	No. of clicks	No. of clicks
Can we correct the source term if it hasn't been extracted correctly?	yes/no	yes/no
How easy/difficult is it to correct the source term?	No. of clicks	No. of clicks
Can we correct the target term if it hasn't been extracted correctly?	yes/no	yes/no
How easy/difficult is it to correct the target term?	No. of clicks	No. of clicks

Table 5: Tests for maintainability

Performance efficiency

For all the previous tests we will be using a small corpus, but we think that it might be interesting to analyse how the tools react to a larger corpus. We will be testing the **time behaviour** of both tools when performing the extraction of a large corpus compared to processing a small corpus and doing a manual extraction. As large corpus we will use the complete *Guide to Meteorological Instruments and Methods of Observation* (WMO-No. 8) with its translation which accounts for 390, 855 words in English and 458, 791 words in Spanish.

Feature	MultiTrans Translation Generator	SynchroTerm
How much time does it take to process a small corpus?	min:sec	min:sec
How much time does it take to process a large corpus?	min:sec	min:sec

Table 6: Tests for performance efficiency

Why we are not testing reliability

Another very interesting characteristic is **reliability**, with subcharacteristics **maturity**, **availability**, **fault tolerance**, and **recoverability**. **Fault tolerance** and **recoverability** are particularly interesting to test because when extracting terminology we are usually dealing with large amounts of data and it can cause systems to crash. If the systems do crash, it is interesting to evaluate if they are capable of recovering the information we were working on (for example, if the candidate list is recoverable with the changes that had been introduced or if we have to start all over and perform the extraction again). These types of tests, forcing systems to crash for example, is more related to developers of software than end-users, and fall outside the scope of this study.

Numerical values

In all these tests, when the answer to a question is 'yes', the tool will be awarded 10 points, and when the answer is 'no', the tool will be awarded a 0.

For questions where we measure the number of clicks, the tool that requires fewer clicks will be awarded 10 points, and the tool that requires more clicks will get a 0. For example, if MultiTrans requires 10 clicks to modify the exclude list and SynchroTerm requires 5, MultiTrans will get a 0 and SynchroTerm will get a 10.

Where the measure is time, the less time the tool takes to process the corpus the better, so 10 points will be awarded to the tool that takes less time and a 0 will be awarded to the tool that takes more time. Manual extraction will definitely take more time than either software, so we will keep this measurement for future reflection and not for arriving to a numerical total.

In order to give a numerical value to precision and recall, we first want to remember what we said in Chapter 3, section 3.1, when we agreed with (L'Homme, 2004) that the best option was a term extractor that reduces silence even if it increases noise since it is easier to clean a list than to find a missing term in a corpus. Silence is low when recall is high, this is why recall will be considered more important than precision. The tool that has the better recall will get 50 points, and the tool with the worst recall will get a 0. As for precision, the tool that has the better precision will get 20 points and the other one will get a 0.

Once the tests are done, all results will be added, and the tool with the higher score is the tool that fulfils most of LSU requirements.

Characte-ristic	Subcharacte-ristic	Feature	Possible answers	Points
Functional suitability	Functional completeness	1) Can we do a bilingual extraction with the software?	Yes No	10 0
		2) Does the software support pdf files?	Yes No	10 0
		3) Does the software support word files (doc and docx)?	Yes No	10 0
		4) Does the software support all six UN languages (Arabic, Chinese, English, French, Russian, and Spanish)?	Yes No	10 0
		5) Does it support French as interface language?	Yes No	10 0
		6) Can we see the extracted term in context?	Yes No	10 0
		7) Does the software provide a source for each extracted term?	Yes No	10 0
		8) Can the software process tables and images in the corpus?	Yes No	10 0
		9) Is the software capable of extracting terminology from tables?	Yes No	10 0

	Functional correctness	10) Recall	higher number lower number	30 0
		11) Precision	higher number lower number	20 0
Compatibility		12) Does the tool export results in Excel?	Yes No	10 0
		13) Does the tool export results in an international exchange format (tbx, xml, html)?	Yes No	10 0
Maintainability	Modifiability	14) Can we customize the software (does it have an ignore list)?	Yes No	10 0
		15) Can we edit the ignore list?	Yes No	10 0
		16) How easy difficult is it to edit the ignore list?	fewer clicks more clicks	10 0
		17) Can we correct the source term if it hasn't been extracted correctly?	Yes No	10 0
		18) How easy/difficult is it to correct the source term?	fewer clicks more clicks	10 0
		19) Can we correct the target term if it hasn't been extracted correctly?	Yes No	10 0
		20) How easy/difficult is it to correct the target term?	fewer clicks more clicks	10 0
Performance efficiency	Time behaviour	21) How much time does it take to process a small corpus?	less time more time	10 0
		22) How much time does it take to process a large corpus?	less time more time	10 0

Table 7: Numerical values for all tests

In Table 7 above we have regrouped all the features that will be evaluated with the characteristics of ISO/IEC 25010 to which they correspond. We have also included the possible answers to our questions and the points each tool would get according to the answer.

6.1 Execution of the evaluation

In this section we are going to explain how the evaluation will be carried out.

For feature inspection tests, the evaluator will perform all the tests and answer the yes/no questions. The number of clicks will be measured manually and time will be measured with the built-in stopwatch of an iPhone 4. The evaluator will have a hard copy of Table 7 and will answer the questions as she performs the tests.

To measure precision and recall will be slightly more complicated than that. In section 3.1 we quoted three techniques to assess the performance of term extractors from (L'Homme, 2004). We will be using the third technique, which consists of building a reference list and then comparing the candidate list that the term extractor produces with this list, to be able to calculate the noise and silence that each tool produces.

In order to build our reference list, we want to use TermoStat to avoid performing a manual extraction of our corpus. To verify that we can use TermoStat, we will perform a short test. We will select part of our corpus and perform a manual extraction. Even though the results of manual extractions may vary according to who performs the extraction, it is generally believed that they are of the best quality possible (certainly, there is no noise, though there might be some silence). The manual extraction will be done by the evaluator. Then, we will submit this same part of the corpus that what used for manual to extraction to TermoStat and we will hopefully obtain as result a list of candidate terms with the same terms as those of the manual extraction. This will mean that TermoStat is usable to create a reference list and we can move forward with our tests.

To assess the performance of our tools we will first submit the remaining part of the corpus to TermoStat. The result of this extraction will constitute our reference list against which we will compare the extraction results of our tools.

Then, we will submit the same remaining part to both MultiTrans and SynchroTerm. We will compare the extraction results to our reference list and calculate precision and recall for both tools.

Before being able to carry out these tests, there are several arrangements to be made, like dividing the corpus, which we will explain in the next section.

6.2 Pre-test arrangements

We have already defined what we are going to measure and how we are going to do it. Now, before actually performing the tests, we have to prepare the corpus for the extraction.

For the first publication mentioned in Chapter 2, *Abridged Final Report with Resolutions and Recommendations of the Extraordinary Session (2014) of the Commission for Basic Systems* (WMO-No. 1140), we obtained two Word files from the English and Spanish editors at WMO, one for each language. The first thing we had to do was to extract the sections of this publication that we are going to use for our tests, that is the resolutions and recommendations, without their respective annexes, which are the only sections that are revised, edited and reliable for terminological extraction. We were left with 8,834 words in English and 11,253 words in Spanish. We carefully compared these files side by side to make sure that that they had the same amount of paragraphs and that nothing was missing from the original to the translation for alignment to be perfect.

As for the second publication, *Guide to Meteorological Instruments and Methods of Observation* (WMO-No. 8), we received three Word files for the English version (Parts I, II and II) from the English editors at WMO and for the Spanish version we received 39 files, one per chapter, from AEMET[17]. We copied into new files the section we are going to use for the tests, that is Chapter 11 of Part II. We checked these files side by side to make sure that they had the same amount of sections and paragraphs. This took some editing, especially for images. Some items of the English file were introduced in the Spanish version as a table with two columns 'original' and 'translation' with an indication of what item of the original they were translating for the designer to introduce the changes prior to publication. These tables and indications had to be deleted since they constituted a

[17] The Spanish State Meteorological Agency, as explained in Chapter 2.

difference between the original and the translation which would probably translate as an error in the alignment of both texts. Another part that was deleted was the bibliography. This publication offers a bibliography after every Chapter, and we believe that this information is not relevant in a terminological extraction, so we deleted it in both languages. The rest was untouched. We were left with 15,217 words in English and 17,629 words in Spanish.

The tools that we are testing will be performing bilingual extractions so the first thing they have to do is to align original and translation. This is why we have checked everything for alignment to be perfect. We will not be testing the quality of the alignment in these tools.

As mentioned in the previous section, we have to select a part of the corpus for manual extraction in order to establish a reference list. We also want to save part of the corpus as unseen data (i.e. without submitting it to the extractors) in case we have to do further tests. This means that we have to create sub-corpora and we will call them C_1, C_2 and C_3, where C_1 is the sub-corpus for manual extraction, C_2 is the sub-corpus of unseen data and C_3 is the sub-corpus for the tests. We will do this for each publication, so we will end up with C_{1a}, C_{2a} and C_{3a} for publication WMO-No. 1140, and C_{1b}, C_{2b} and C_{2b} for publication WMO-No. 8. We have decided to divide the corpus arbitrarily and to assign the first 15% (number of words) of each publication to C_1, the second 15% of each publication to C_2 and the remaining 70% to C_3. See recapitulation chart below.

	Sub-corpora for WMO-No. 1140	**Sub-corpora for WMO-No. 8**	**% of words from the complete Corpus**
Manual extraction	C_{1a}	C_{1b}	15%
Unseen data	C_{2a}	C_{2b}	15%
Automatic extraction	C_{3a}	C_{3b}	70%

Table 8: Recap chart for corpus division

When dividing the corpus, we realised that strictly respecting these percentages meant cutting sentences or sometimes paragraphs in half. We tried to avoid this and to respect the percentages as closely as possible. The results are as follows:

Sub-corpora for WMO-No. 1140	**% of words from the complete Corpus**	**Actual No. of words**
C_{1a}	15% (1,325 words)	1,325
C_{2a}	15% (1,325 words)	1,312
C_{3a}	70% (6,183.8 words)	6,198

Table 9: Sub-corpora for publication WMO-No. 1140

Sub-corpora for WMO-No. 8	**% of words from the complete Corpus**	**Actual No. of words**
C_{1b}	15% (2,282.55 words)	2,241
C_{2b}	15% (2,282.55 words)	2,239
C_{3b}	70% (10,651.9)	10,737

Table 10: Sub-corpora for publication WMO-No. 8

Another arrangement before performing the tests is the creation of an extraction record. If we recall the scheme of the terminographic work that we proposed in Chapter 2, in stage 3.1, *Gathering information on terms*, we mentioned that we will be using an extraction record when we performed manual extraction. Extraction records are used to collect information on terms and enable us to see how the terms behave within the corpus and their frequency. Our extraction records have been done in Microsoft Excel and below you can see the model we have used:

Extraction record for publication WMO-No.					
Person performing manual extraction:					
Time it took to perform: (hh:mm:sec)					
Term in English	Acronym English	Extra info if pertinent	Term in Spanish	Acronym Spanish	Extra info if pertinent

Table 11: Extraction record to be used in manual extraction

From that extraction record, a terminological record was created for each term. The terminological record has also been done on Microsoft Excel, with the following model:

Terminological records								
TermEN	Source TermEN	Synonym1 TermEN	Synonym2 TermEN	TermES	Source TermES	Synonym1 TermES	Synonym2 TermES	Note

Table 12: Terminological record to be used in manual extraction

Once all of this was established, we moved forward with our short test to assess if we can use TermoStat to build our reference list.

Manual extraction has been done using a hard copy of sub corpora C_{1a} and C_{1b} because of personal preference of the evaluator. Every time a term appeared it was highlighted in the hard copy and then inserted into the Excel file. The manual extraction was performed on the English version of the sub corpora and then the Spanish equivalents were searched for and inserted into the extraction record. It is important to take into account the differences between the publications used because of the different terms that they will produce. One publication is a session report and the other is a technical one, so we came across the following different types of terms:

- Titles of publications
- Names of programs, working groups, task teams
- Names of WMO bodies
- Technical terms
- Instruments used to perform meteorological observations and parts of instruments

- Methods and measurement procedures

- Names of variables that are measured

In the cases where we were not sure if we were dealing with a term or not we performed a research in METEOTERM and in other WMO documents and publications.

We were also faced with terms that seemed as they belonged to the common language, but in our corpus they have a special significance. For example, the term "local scale" does not seem like a term, but in our text it has special significance, that is why we keep it as term. Extract from corpus:

> "(…) There is no more important an input to the success of an urban station than an appreciation of the concept of scale. There are three scales of interest (Oke, 1984; Figure 11.1): (…)
>
> (b) Local scale: This is the scale that standard climate stations are designed to monitor. It includes landscape features such as topography, but excludes microscale effects. In urban areas this translates to mean the climate of neighbourhoods with similar types of urban development (surface cover, size and spacing of buildings, activity). The signal is the integration of a characteristic mix of microclimatic effects arising from the source area in the vicinity of the site. The source area is the portion of the surface upstream that contributes the main properties of the flux or meteorological concentration being measured (Schmid, 2002). Typical scales are one to several kilometres; …"

All the variations of the terms appear in the extraction record, but not on the terminological records. These variations will not be taken into account when evaluating the extraction tools (for example, if a term usually carries a hyphen and the tools offer the term as a candidate without the hyphen, the candidate will be considered a valid term. The same will apply for plurals and singulars: the candidate term will be considered a term regardless if it was extracted by the tool in the plural or the singular form).

For complex terms, if the tools extract only part of them, they will be considered valid terms since they can be modified thanks to the context offered by the tools and edited once imported into a termbase.

Once the manual extraction was done and a reference list established, we compared the results of TermoStat extraction with our list. We realised two things while doing this. The first one was that while performing the manual extraction we omitted 6 terms. The second one was that we will not be able to use TermoStat extraction results exactly as the extractor provides them. Silence was rather high for one of the publications and noise was high for both publications. See the table below.

	TOTAL TERMCOUNT on reference list	Candidates Termostat extracted	Terms Termostat extracted	Precision	Noise	Recall	Silence
Corpus C_{1a} - WMO-No. 1140	53	101	22	0,23	77%	0,42	58%
Corpus C_{1b} - WMO-No. 8	86	257	59	0,23	77%	0,69	31%

Table 13: Precision and recall of TermoStat for sub-corpora C_{1a} and C_{1b}

With these results, we decided that the reference list for the larger part of the corpus should be done manually with the help of TermoStat, to make sure that no terms were omitted and in an attempt to avoid subjectivity. For the manual extraction of sub-corpora C_{3a} and C_{3b} we used the same procedure as the one explained for sub-corpora C_{1a} and C_{1b}.

In the following section we will describe how we executed our tests. We will list all the features, answer the questions of our tests and give explanations where needed.

6.3 The tests

As explained in the previous section, the evaluator performed the tests with a hard copy of Table 7 where she wrote down the answers to the questions as she performed the tests. The tool that was first tested was MultiTrans at WMO, and then SynchroTerm which was installed on a notebook. The notes that were taken down during the tests are included below.

Next we are going to answer the questions of our tests.

Feature 1: Can we do a bilingual extraction with the software?

Yes, both tools can perform bilingual extractions.

Feature 2: Does the software support pdf files?

Yes, both tools can perform a bilingual extraction from a pdf file.

Feature 3: Does the software support word files (doc and docx)?

Yes, both tools were able to perform an extraction from both types of files.

Feature 4: Does the software support all six UN languages (Arabic, Chinese, English, French, Russian, and Spanish)?

Yes, both tools support all six languages.

Feature 5: Does it support French as interface language?

Both tools support French as an interface language.

Feature 6: Can we see the extracted term in context?

Both tools offer the context of the term as soon as the term is selected.

Feature 7: Does the software provide a source for each extracted term?

In MultiTrans translator generator, when importing into the working termbase, the second Spanish term is imported without source, reason why it was given a 0 (see Figure 6 below). When the termbase is exported, the source appears only for the first term.

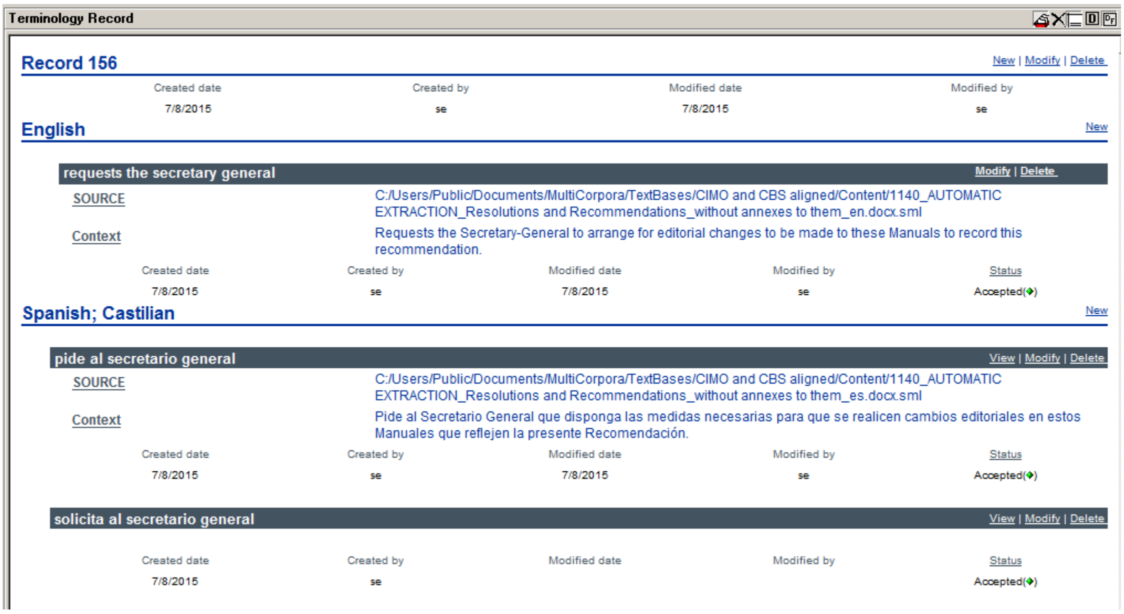

Figure 6: MultiTrans METEOTERM desktop version – Spanish synonym is inserted without source from translation generator.

In SynchroTerm, terms that are validated appear with a source in the termbase that the tool creates, but when the termbase is exported, the source does not appear in the export file. That is why it was given a 0 (see Figures 7 and 8 below).

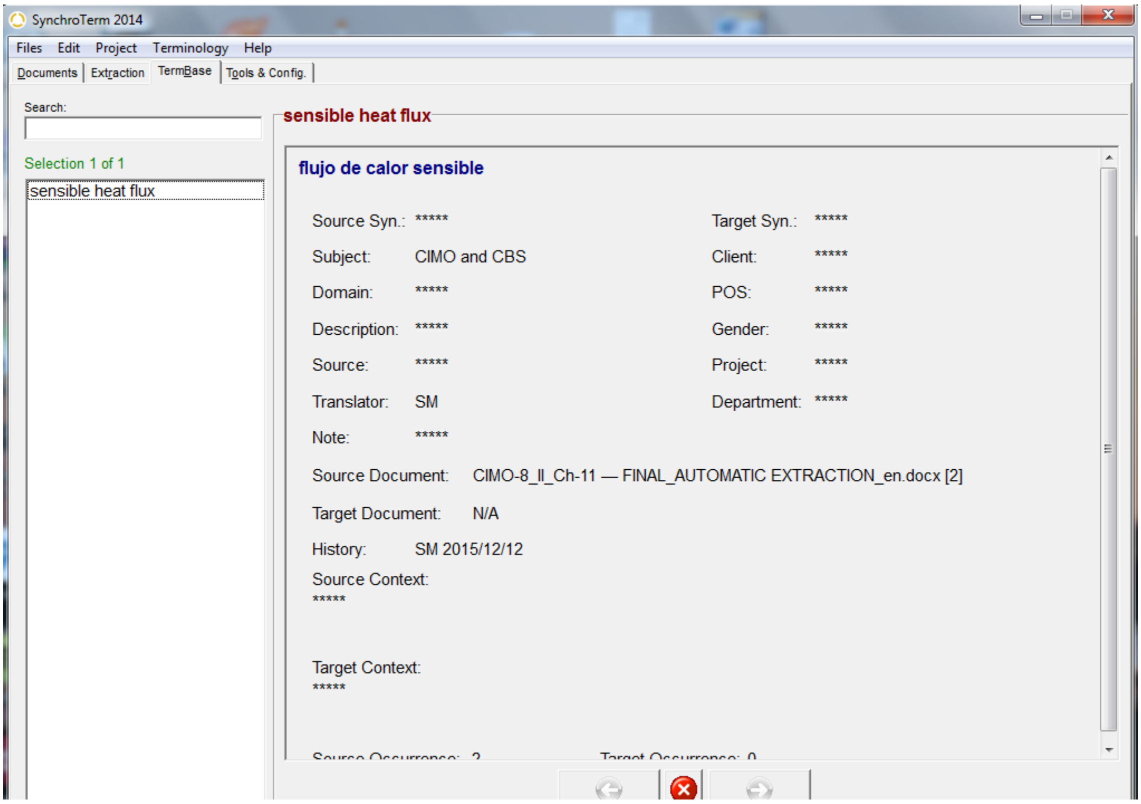

Figure 7: SynchroTerm shows source for terms in the termbase used with each extraction project.

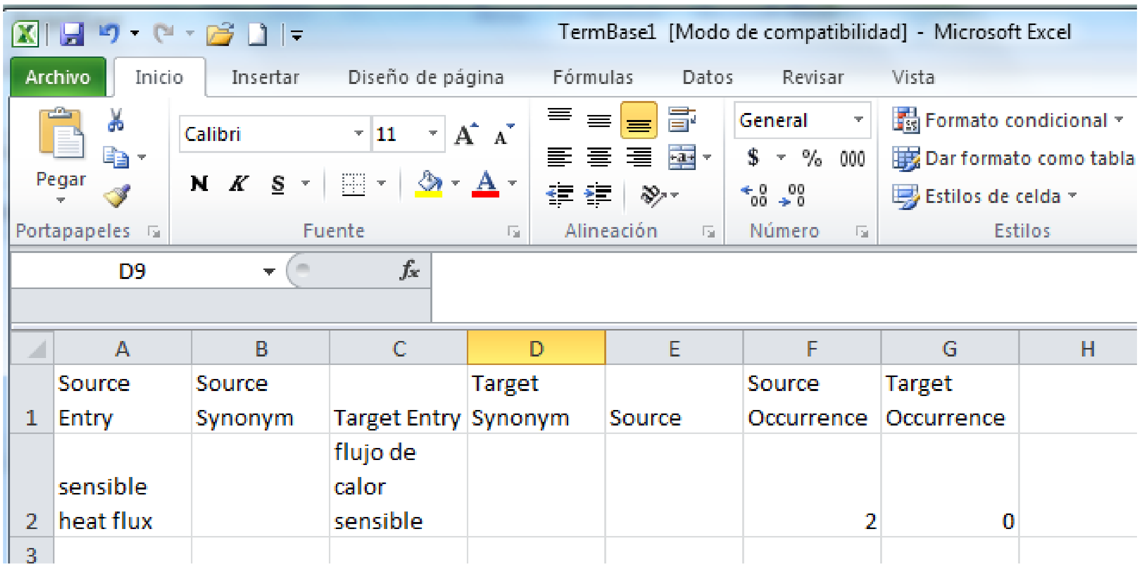

Figure 8: SynchroTerm does not show the source for terms exported.

Feature 8: Can the software process tables and images in the corpus?

Neither tool crashed when dealing with the corpus, so they both can process tables and images.

Feature 9: Is the software capable of extracting terminology from tables?

There was only one table in the corpus and unfortunately it did not contain any terms. Since it was a draw both tools were awarded a 0.

Features 10 and 11: Recall and precision

In the previous section we clarified that candidate terms would be considered terms regardless of whether the tool extracted them with or without a hyphen and regardless of the variations. We also came across another difficult case, when candidate terms were 'repeated' in the candidate term list. For example, in our reference list we have the term 'instrument enclosure'. Synchroterm offers as candidates the terms 'enclosure' and 'instrument enclosure'. In this case the latter was considered a term. (Though if the tool hadn't offered the second candidate, the first one would have been considered a valid term, since complex terms that are partially extracted are also considered terms).

We were presented with another type of difficulty for simple terms. Sometimes simple terms were extracted by tools within a larger structure. For example, in our reference list we have the acronym 'SATCOM', and the only candidate term that Synchroterm offers that contains 'SATCOM' is 'satcom forum'. This candidate will be considered as a term, since it can be edited when it is validated. (MultiTrans offers 'satcom forum' and 'satcom meetings' as candidates, only one was validated as a term).

A positive surprise was the fact that both tools extracted the names of publications, working groups and task teams. Even if they did not extract all of them and some of them partially, it was a surprising result.

One of the regular settings for term extractors is to set minimum frequency at 2. This leaves all terms that appear only once undetected. A test was done with SynchroTerm to assess if changing this parameter would change the results significantly, and it was not the case. Noise was increased substantially (from 71% to 84%, which means the terminographer would need more time to go through the candidate term list validating terms, and recall was not highly increased –from 0,32 to 0,37, which means that silence remains over 60%). Our conclusion is that changing the settings will not increase recall significantly, and recall is the most important measure in our study, therefore the numbers used in this study are from performing the extractions with minimum frequency set at 2.

In Table 15 below we present the results of the tests for precision and recall for MultiTrans:

	TOTAL TERMCOUNT on reference list	Candidates MultiTrans extracted	Terms MultiTrans extracted	Precision	Noise	Recall	Silence
Corpus C_{3a} and C_{3b} (WMO-No. 1140 + WMO-No. 8)	405	416	163	39%	61%	40%	60%

Table 14: Precision and recall of MultiTrans for sub-corpora C_{3a} and C_{3b}

From this table we can see that from the candidate list that MuliTrans offered only 39% of those candidates were actually terms (precision). Of all the terms that were present in the corpus, MultiTrans extracted only 40% of them (recall), this means that 60% of the terms were not detected by the tool (silence).

Table 16 shows the results for SynchroTerm:

	TOTAL TERMCOUNT on reference list	Candidates Synchro-Term extracted	Terms Synchro-Term extracted	Precision	Noise	Recall	Silence
Corpus C_{3a} and C_{3b} (WMO-No. 1140 + WMO-No. 8)	405	443	130	29%	71%	32%	68%

Table 15: Precision and recall of SynchroTerm for sub-corpora C_{3a} and C_{3b}

From this table we can see that from the candidate list that SynchroTerm offered only 29% of those candidates were actually terms (precision). Of all the terms that were present in the corpus, SynchroTerm extracted only 32% of them (recall), this means that 68% of the terms were not detected by the tool (silence).

Feature 12: Does the tool export results in Excel?

Yes, both tools export results in Excel (see next question for further explanation on exporting options for both tools).

Feature 13: Does the tool export results in an international exchange format (tbx, xml, html)?

Both tools work similarly. The list of candidate terms can only be exported in .txt format in both tools. In order to export terms and translations, we first have to validate them and then export them into the working termbase. From the termbase we can export into different formats. From MultiTrans we can export into .txt, Microsoft Excel, TBX, XML and HTML. From SynchroTerm we can export into HTML, Trados MultiTerm 5.5, Terminotix LogiTerm, Comma delimited (.csv), Trados MultiTerm iX, Microsoft Excel, Trados WinAlign and PROMT.

Feature 14: Can we customize the software (does it have an ignore list)?

Both tools provide a default ignore list which can be edited.

One important difference in customization between the tools is the fact that SynchroTerm can be set to extract only nouns. *A priori*, it seemed that this would mean that noise would be lower for SynchroTerm. Still, noise was higher in SynchroTerm than in MultiTrans (71% vs 61%).

Another interesting customization option that SynchroTerm offers is that the user can filter out terms from the extraction based on a list created by the user. If the user has a unilingual list of terms for which it needs to find the translation in a given corpus, the Restrictive Terminology option will find the correspondances of the terms contained in that list only.

Feature 15: Can we edit the ignore list?

Yes, both tools allow for the ignore list to be edited by the user.

Feature 16: How easy difficult is it to edit the ignore list?

To edit the ignore list in MultiTrans and the stop list in SynchroTerm, we just have to open the corresponding .txt file and edit it, so we just need 4 clicks. Since it is a draw, both tools were awarded a 0.

Feature 17: Can we correct the source term if it hasn't been extracted correctly?

In MultiTrans translator generator we cannot modify the English term in the extraction list before exporting it to the working termbase; we can only validate or discard the extracted terms.

In SynchroTerm we can modify the English term prior to exporting it to a termbase. MultiTrans was awarded a 0 and SynchroTerm was awarded a 10.

Feature 18: How easy/difficult is it to correct the source term?

This question cannot be attributed any points. Since MultiTrans does not allow the modification of the source term, it is no longer applicable. (For information, in SynchroTerm it just takes one click for selecting the correct English term from the context to get the correct source term).

Feature 19: Can we correct the target term if it hasn't been extracted correctly?

Both tools allow the user to correct the target term before validation.

Feature 20: How easy/difficult is it to correct the target term?

In MultiTrans we double click on the proposed target term that we want to modify and the field is activated for us to type in it. In SynchroTerm we just click once to select the appropriate part of context that we want to be considered the target term. SynchroTerm was awarded a 10 and MuliTrans a 0.

Feature 21: How much time does it take to process a small corpus?

MultiTrans translator generator performs the extraction in two steps: firstly it extracts the terminology in English and then it generates the translations. We have arrived to the total by summing up the time it has taken the software to complete the two necessary steps: 00:22 to extract terminology and 01:07 to generate the translations. It has taken this tool 01:29 to process a small corpus.

SynchroTerm performs the extraction in one step. It has taken this tool 00:19 to process a small corpus.

Feature 22: How much time does it take to process a large corpus?

The comments of the previous question apply to the large corpus, and the times for both steps in MultiTrans for a large corpus were 00:34 to extract the terminology and 46:55 to generate the translations, totalling **47:29**. In the case of SynchroTerm, it has taken **07:33**.

Only for information, the amount of candidate terms that we obtained for each tool when they extracted terminology from this large corpus was 11, 187 for MultiTrans and 7, 904 for SynchroTerm.

It is interesting to note the time it has taken both tools to upload the corpus from which they extracted the terminology. For a small corpus it has taken

MultiTrans 00:45 seconds and 07:42 for a large corpus, though these times have not been added to the final result.

The times for uploading the corpus in SynchroTerm were 00:03 for a small corpus and 00:12 for a large corpus.

With regards to time it is worth noting that the results obtained in the evaluation are to test the speed of the tools. The whole terminographic work does not finish there, since LSU officers then have to assess the candidate list to validate the terms and then select the correct translations. After that they would have to revise the work (with Spanish revisers and experts if needed) and finally upload the terminological records to METEOTERM.

Performing the manual extraction of the corpus has taken the evaluator 25 hours. The question that arises is the following: Does the use of a tool reduce the time a terminographer spends on a given terminographic project? Unfortunately, there is no answer to this question. Evaluating whether a tool actually translates into a gain of time for the terminographer falls out of the scope of the research, since we are only evaluating tools, but it would be an interesting test to perform.

In this section we have answered all the questions of our tests. In the next section we give the numerical results.

6.4 Test results

We present the following table with the results of our tests for a clear visualization:

Characte-ristic	Subcharacte-ristic	Feature	MultiTrans Translation Generator	Synchro Term
Functional suitability	Functional completeness	1) Can we do a bilingual extraction with the software?	10	10
		2) Does the software support pdf files?	10	10
		3) Does the software support word files (doc and docx)?	10	10
		4) Does the software support all six UN languages (Arabic, Chinese, English, French, Russian, and Spanish)?	10	10
		5) Does it support French as interface language?	10	10

		6) Can we see the extracted term in context?	10	10
		7) Does the software provide a source for each extracted term?	0	0
		8) Can the software process tables and images in the corpus?	10	10
		9) Is the software capable of extracting terminology from tables?	0	0
	Functional correctness	10) Recall	50	0
		11) Precision	20	0
Compatibility		12) Does the tool export results in Excel?	10	10
		13) Does the tool export results in an international exchange format (tbx, xml, html)?	10	10
Maintainability	Modifiability	14) Can we customize the software (does it have an ignore list)?	10	10
		15) Can we edit the ignore list?	10	10
		16) How easy difficult is it to edit the ignore list?	0	0
		17) Can we correct the source term if it hasn't been extracted correctly?	0	10
		18) How easy/difficult is it to correct the source term?	-	-
		19) Can we correct the target term if it hasn't been extracted correctly?	10	10
		20) How easy/difficult is it to correct the target term?	0	10
Performance efficiency	Time behaviour	21) How much time does it take to process a small corpus?	0	10
		22) How much time does it take to process a large corpus?	0	10
TOTAL			**190**	**160**

Table 16: Test results

As we can see from this table, the tool that has obtained the higher score is MultiTrans.

Even though these scores seem high, let us not jump into conclusions just yet. If we take a closer look at the table we can see that both tools offer the same functionalities, only in SynchroTerm they are sometimes easier to use (for example, in SynchroTerm the source term can be modified either at the moment of validating the term or once it has been imported into the working termbase, whereas in MultiTrans it can only be modified once it has been imported to the working termbase, but it can still be modified).

One problem that both tools present is with regards to the sources of the terms. The best solution we can think of is to submit to extraction one publication at a time, this way, we can be sure that the source for all terms is the same.

We have noticed a great difference regarding the time behaviour of both tools when processing a large corpus. This is probably due to the fact that MultiTrans at WMO is server-based and it probably slows the process down (it should be noted that the test was performed on a Saturday in an attempt to avoid a large number of users using the server). It would be interesting to re-test, out of curiosity, the speed of both tools using the same computer and avoiding server-based connections. In practical terms we do not think that working with corpora as large as the one we have used for the test is advisable: going through a list of 11, 187 terms would be a task of several months. We strongly believe that the best option would be to divide a corpus that large into smaller parts to make the task more bearable.

As we have explained before, the most important values when evaluating term extractors are precision and recall. We are rather surprised by the results for two reasons. The first one is that we expected SynchroTerm to present better results than MultiTrans; the second one is that we expected better results from both tools.

Unfortunately, there has not been a great difference between the tools as predicted at the beginning of this Chapter when we discussed functional suitability. There is a 10% difference in precision (39% for MultiTrans and 29% for SynchroTerm) and an 8% difference in recall (40% for MultiTrans and 32% for SynchroTerm). Both tools failed to detect at least 60% (and 68% for SynchroTerm) of the terms that were present in the corpus. We can see that there are still a lot of improvements to be made to term extractors and that these tools could be used as support by terminographers, but they are far from producing the same results as humans. Hopefully technological advances will enable us to get there in the years to come.

7 Conclusion

The research work that has been carried out has two aspects to it. One of them is the terminological aspect, and the other is the CAT tool evaluation aspect.

For the terminological aspect, basic concepts of terminology have been introduced in the first chapters. The methodologies proposed by renowned authors in the field have been followed and we believe that the terminographic work that has been done is accurate.

The CAT tool evaluation aspect has been developed following international standards and an evaluation framework that has had a great impact on the field. The standard and the framework have been adapted to a specific type of CAT tool: term extractors.

Both these aspects have been placed in context: the World Meteorological Organization.

The purpose of this research was to find the term extractor that best satisfied the needs of LSU officers within the Language, Conference and Publishing Services Department of WMO, and the one that would enable them to extract terms from publications as automatically as possible. The only terminological extraction that currently takes place at WMO is manual extraction of session reports (in search of, in particular, names of special groups, task teams, panels, commissions that are created and other institutional terminology).

The tests have been carried out and as we can see from Table 16 in the previous Chapter, the tool that has obtained the highest score, and therefore fulfils most of LSU officers' requirements, is MultiTrans. This is very positive, since it is the tool that is already in place at WMO. Nevertheless, this result does not imply that LSU officers can start performing term extractions with this tool and will be satisfied with the results. Actually, the results have been very disappointing for both tools. They both present a very low percentage of precision (the proportion of terms that are actually terms of the candidate list): 29% for SynchroTerm and 39% for MultiTrans. Precision is important, but we established beforehand that most important to us was recall (the proportion of terms extracted of all the terms appearing in the corpus). MultiTrans had a recall of 40%, which means that 60% of the terms went undetected. SynchroTerm showed worse results with recall of 32%, which means that 68% of the terms went undetected.

Leaving 60% of terms behind when we are perfoming a terminological extraction is far from an ideal scenario, but leaving them behind in the extraction of a session report is not an option. The names of teams, groups, panels and commissions are needed for future translations, especially for translating correspondence, which represents a large volume of everyday translations. We cannot recommend LSU officers to use MultiTrans to perfom their term extractions of session reports even if it obtained the highest score. 60% of silence is just too high. Our recommendation is that they continue with manual extraction of session reports. Every term present in these types of documents is too valuable to remain undetected. We could even typify these terms as 'urgent', since all correspondence that is exchanged after the report is published needs to contain the correct terms and their translations.

However, for technical publications, such as WMO-No.8, MultiTrans could be used for extracting terms. Maybe if we used larger parts of these types of publications (instead of one chapter as was done in this evaluation), we could increase the frequency of some terms, thus improving the recall value of the tool. The terms present in these publications, unlike the ones present in session reports, though important, they are not 'urgent'. We believe that adding 40% of the terms present in a given publication into METEOTERM would be better than making no additions. With the current resources available within LSU and taking into account all the different tasks that they have to perform, there is definitely no time for an LSU officer to read an entire technical publication and extract the terms. MultiTrans offers the opportunity to add 40% of terms present in a given publication into the terminological database without having to read it. It could be a fair compromise.

We can conclude that the results of this evaluation show that one tool is slightly better than the other one, but unfortunately they do not establish that any of the two tested tools is good at extracting terms. Our recommendations for WMO are to continue with the current procedure for the extraction of terminology of session reports (manual extraction) and to take advantage of the percentage of terms that MultiTrans enables them to add to METEOTERM, which, though it is far from perfect, is better than no terms at all.

Where do we go from here?

We have evaluated two tools for their extraction in English and Spanish. It would be interesting to repeat this evaluation with other language combinations to verify if the results for precision and recall are the same when extracting from a different source language. We would also like to evaluate how MultiTrans Analysis Agent performs, since we were unable to do so for technical issues.

A hypothesis presented during our conclusion was if MultiTrans could improve its recall value by using a larger portion of a technical publication as extraction corpus. Using a larger corpus could increase the frequency of some terms, making them detectable to MultiTrans. It would be interesting to repeat the tests using a larger portion of a technical publication as extraction corpus to verify this hypothesis.

During our research we were tempted to include the time variable several times, but that would have meant surpassing the scope of this research. Apart from the time a tool takes to extract terminology, which will always be less than the amount of time it takes a human to perform a terminological extraction, it would be interesting to evaluate if the use of tools actually translates into a gain of time for the terminographer taking into account all the steps in the terminographic work (established in Chapter 2). This has not been done yet, probably because of the amount of resources needed, but such an evaluation would certainly be an important one in the field.

Bibliography

Benavent, P., & Parrilla, S. (n.d.). Análisis de la extracción automática de términos con el programa informático extraterm. *Jornades de Foment de La Investigació*. Retrieved from http://www.uji.es/bin/publ/edicions/jfi12/25.pdf

Bernier-Colborne, G. (2012). Élaboration D'Un Corpus Étalon Pour L'Évaluation D'Extracteurs De Termes, 155.

Bourigault, D., Jacquemin, C., & L'Homme, M.-C. (Eds.). (2001). *Recent Advances in Computational Terminology*. John Benjamins Publishing. Retrieved from http://books.google.com/books?hl=en&lr=&id=0eX8HtA7mQIC&pgis=1

Bowker, L. (2002). *Computer-aided Translation Technology: A Practical Introduction*. University of Ottawa Press. Retrieved from https://books.google.com/books?hl=en&lr=&id=ly29-mc6dO0C&pgis=1

Bowker, L., & Pearson, J. (2002). *Working with Specialized Language: A Practical Guide to Using Corpora*. Routledge. Retrieved from https://books.google.com/books?id=RqqBAgAAQBAJ&pgis=1

Cabré, M. T. (1999). *Terminology: Theory, Methods, and Applications*. Retrieved from http://books.google.ch/books/about/Terminology.html?id=GAqGD9Xtu0IC&pgis=1

Cabré, M. T., Estopà Bagot, R., & Vivaldi Palatresi, J. (2001). Automatic Term Detection: a review of current systems. In *Recent Advances in Computational Terminology* (pp. 53–88). Retrieved from http://www.upf.edu/pdi/dtf/rosa.estopa/docums/ract_fin.pdf

Drouin, P. (2003). Term extraction using non-technical corpora as a point of leverage. *Terminology*, *9*(1), 99–115. doi:10.1075/term.9.1.06dro

EAGLES. (1996). Evaluation of Natural Language Processing Systems, *5*. Retrieved from http://www.issco.unige.ch/en/research/projects/ewg96/ewg96.html

EAGLES. (1999). The EAGLES 7-step recipe. Retrieved September 3, 2015, from http://www.issco.unige.ch/en/research/projects/eagles/ewg99/7steps.html

Estopà Bagot, R. (1999). Eficiencia en la extracción automática de terminología. *Perspectives*, *7*(2), 277–286. doi:10.1080/0907676X.1999.9961365

Gervais, D., Blvd, S., & Jy, C. (n.d.). MultiTrans ™ System Presentation Translation Support and Language Management Solutions. *Framework*.

Hutchins, J. (1996). From the Archives... ALPAC: the (in)famous report, (14), 9–12.

ISO/IEC 25010. (2011). ISO/IEC 25010:2011 - Systems and software engineering -- Systems and software Quality Requirements and Evaluation (SQuaRE) -- System and software quality models. Retrieved September 2, 2015, from http://www.iso.org/iso/catalogue_detail.htm?csnumber=35733

Jacquemin, C. (2001). *Spotting and Discovering Terms Through Natural Language Processing*. Retrieved from https://books.google.ch/books/about/Spotting_and_Discovering_Terms_Through_N.html?id=W6AB06SBAGMC&pgis=1

Kageura, K., & Umino, B. (1996). Methods of automatic term recognition: A review. *Terminology*, *3*(1996), 259–289. doi:10.1075/term.3.2.03kag

King, M. (1992). L'évaluation des systèmes de traduction automatique dans le cadre d'un service de traduction. *Meta: Journal Des Traducteurs*, *37*(4), 817. doi:10.7202/003314ar

L'Homme, M.-C. (2001). Nouvelles technologies et recherche terminologique. Techniques d'extraction des données terminologiques et leur impact sur le travail du terminographe, 1–32.

L'Homme, M.-C. (2004). *La terminologie: principes et techniques*. Retrieved from https://books.google.ch/books/about/La_terminologie.html?id=w222vwf6Mo0C&pgis=1
L'Homme, M.-C. (2008). *Initiation à la traductique*. Linguatech. Retrieved from https://books.google.com/books?id=-Oi1PQAACAAJ&pgis=1
Lauriston, A. (1994). Automatic recognition of complex terms: Problems and the TERMINO solution. *Terminology, 1*(1), 147–170. doi:10.1075/term.1.1.11lau
Lauriston, A. (1995). Criteria for measuring term recognition. *Proceedings of the Seventh Conference on European …*, 17–22. doi:10.3115/976973.976977
Quah, C. K. (2006). *Translation and Technology*. Palgrave Macmillan. Retrieved from https://books.google.com/books?id=YXlhQgAACAAJ&pgis=1
Rubio, M., Pastor, V., & Valero, E. (2009). Análisis de la extracción y validación bilingüe de terminología con el programa informático MultiTerm Extract. *Jornades de Foment de La Investigació*. Retrieved from http://www.uji.es/bin/publ/edicions/jfi14/trad/7.pdf
Sauron, V. (2002). Tearing out the Terms: Evaluating Terms Extractors. Retrieved July 5, 2015, from http://www.mt-archive.info/Aslib-2002-Sauron.pdf
TermCoord European Parliament. (n.d.). TermCord Terminology Corrdination Unit. Retrieved July 24, 2015, from http://termcoord.eu/useful-links/free-term-extract

Annex A – WMO authorization to use their publications and to publish results

WMO OMM
World Meteorological Organization
Organisation météorologique mondiale

Secrétariat
7 bis, avenue de la Paix – Case postale 2300 – CH 1211 Genève 2 – Suisse
Tél.: +41 (0) 22 730 81 11 – Fax: +41 (0) 22 730 81 81
wmo@wmo.int – www.wmo.int

TEMPS • CLIMAT • EAU
WEATHER • CLIMATE • WATER

AUTHORIZATION TO PUBLISH INFORMATION

I hereby authorize Ms Sofía Magdalena MEJÍA to conduct the necessary research for her Master's Thesis using WMO publications. I further authorize her to discuss the structure of the Languages, Conference and Publishing Services Department and all matters related to terminology and terminology management in this Department; and to publish the results she has obtained during such research after prior authorization by the Director of Language, Conference and Publishing Services Department.

Done in Geneva, 16 July 2015

Maja Carrieri

Maja Carrieri

Director
Language, Conference and Publishing
Services Department
e-mail: mcarrieri@wmo.int

Annex B – Reference list for sub-corpora C_{1a} and C_{1b}

Reference list manual extraction C_{1a} and C_{1b} WMO-No. 1140	Reference list manual extraction C_{1a} and C_{1b} WMO-No.8
Abridged Final Report with Resolutions of the Sixteenth World Meteorological Congress (WMO-No. 1077)	anemometer
Abridged Final Report with Resolutions of the Sixty-sixth Session of the WMO Executive Council (WMO-No. 1136)	automated station
cache	blending height
CCl	boundary layer
CDMS	calibration
Climate Data Management System	climate observation
Commission for Aeronautical Meteorology	climate station
Commission for Basic Systems	climatic scale
Commission for Climatology	coding
Competency Framework	cosine response
Competency Framework for Public Weather Services Forecasters and Advisers	displacement height
CSB	documentation of metadata
disaster risk	energy conservation
Executive Council	instrument exposure
Executive Council Panel of Experts on Education and Training	fetch
Expert Team on WMO Information System Centres/Task Team on GISCs	first order station
GISCs	flux
Global Framework for Climate Services	forecast
Global Information System Centres	gas analyzer
Guide to Public Weather Services Practices (WMO-No. 834)	heat island
Guide to the WMO Information System (WMO-No. 1061)	height of measurement
hydrometeorological hazard	horizon obstruction

Reference list manual extraction C_{1a} and C_{1b} WMO-No. 1140	Reference list manual extraction C_{1a} and C_{1b} WMO-No.8
hydrometeorological impacts	human activities
impact-based forecasting	hygrometer
Implementation Coordination Team on Information Systems and Services	inertial sublayer
intercommission task team	inertial layer
Manual on the Global Data-processing and Forecasting System (WMO-No. 485)	infrared thermometer
Manual on the WMO Information System (WMO-No 1060)	instrument
National Meteorological and Hydrological Services	internal boundary layer
Open Programme Area Group on Public Weather Services	local scale
partner agencies	mesoscale
Public Weather Services	mast
Public Weather Services forecaster/adviser	meteorological concentration
Regional Association	meteorological environment
risk-based warnings	meteorological observation
Secretariat	meteorological property
Secretary-General	Meteorological Services
Technical Commission	meteorological station
Technical Regulations (WMO-No. 49)	meteorological variable
WIS	microclimate
WIS Core Network	microclimatic effect
WMO constituent bodies	microclimate anomaly
WMO Guidelines on Multi-hazard Impact-based Forecast and Warning Services	microscale
WMO Information System	momentum
WMO Information System (WIS) Network	observing practice
WMO Recommended Practice	perspective circle
World Weather Watch	planetary boundary layer
World Weather Watch Programme for 2012-2015	radiation exchange
	radiometer
	roughness sublayer
	roughness

Reference list manual extraction C_{1a} and C_{1b} WMO-No. 1140	Reference list manual extraction C_{1a} and C_{1b} WMO-No.8
	rural boundary layer
	site selection
	site
	short-wave radiation
	long-wave radiation
	source area
	station
	station site
	surface cover
	thermal turbulence
	mechanical turbulence
	thermodynamic property
	topography
	tower
	turbulent eddies
	turbulent flux
	turbulent transport
	upwelling radiation signal
	urban area
	urban boundary layer
	urban canopy layer
	urban environment
	urban microclimate
	urban observation
	urban station
	waste heat exhaust
	urban terrain
	PBL
	RSL
	RBL
	siting
	footprint
	UBL
	UCL

Annex C – Extraction results by TermoStat of sub-corpora C_{1a} and C_{1b}

Candidates C1a WMO-No. 1140	Candidates C1b WMO-No. 8
cbs-ext	source area
impact-based forecasting	boundary layer
technical commission	fetch
user requirement	microclimate
capacity management	zr
risk-based warning	roughness
global exchange	zh
competence requirement	radiometer
ongoing process	internal boundary
ad hoc intercommission task team	turbulent flux
giscs	zd
competency	sensor
present recommendation	flux
annex	surface type
partner agency	microclimatic effect
hydrometeorological hazard	typical scale
remit of meteorologist	surface cover
provision of public weather	urban station
warning service	climate station
transmission priority	turbulent transport
communication practice	local scale
task team	wind speed
support of social resilience	urban area
top-level competence requirement	blending
knowledge of staff	height
sixteenth session	view factor
requirement review	part i
functional scope	upwind direction
specification publication	internal boundary layer
hydrology	inertial layer
type of environmental datum	blending height
loss of livelihood	scale of interest
adequate uniformity	urban terrain
collaborative mechanism	heat island
overarching approach	fetch requirement
keyword globalexchange	sublayer
abridged final report	vertical
fifteenth session	instrument exposure
ad hoc expert meeting	height of measurement
effective service delivery	urban canopy layer
product exchange	meteorological variable
present resolution	meteorological concentration
wide distribution	footprint

Candidates C1a WMO-No. 1140
competency framework
meteorology
core job-tasks
training of forecaster
regional association
definition of competence requirement
basic forecaster competency
effective operation
capacity requirement
global framework
discovery metadata
necessary resource
related international programme
constituent body
high-priority activity
understanding disaster risk
multilateral collaboration forum
established ongoing exchange of meteorological datum
hydrometeorological impact
specification
cache
forecaster
recommended practice
hydrologist
formal mechanism
information management
destruction of property
requirement
ad hoc
close collaboration
close consultation
recommendation
livelihood
severe weather
climatology
relevant part
datum
work programme
collaboration
capacity
resolution
user
commission
assistance

Candidates C1b WMO-No. 8
surface
layer
air temperature
field of view
waste heat
air flow
scale
surface area
observation
instrument
measurement
radiation exchange
meteorological station
vertical scale
single station
useful and repeatable observation
concept of scale
spatial and temporal scale
special demand
plane surface
airport location
site selection
standard guideline
local internal boundary layer
meaningful observation
inhomogeneity of urban environment
selection of site
unique challenge
standard variable
internal boundary layer height
stable case
long wave radiation
reference level
thermal and mechanical turbulence
actual source area
conceptual representation of source area
standard boundary layer theory
roughness increase
large roughness
zr – zd
displacement height zd
roughness effect
significant site restriction
neutral stability

Candidates C1a WMO-No. 1140	Candidates C1b WMO-No. 8
warning	similar urban terrain
exchange	circular source area
process	mean vertical profile of meteorological variable
ad	mean value of meteorological property
amendment	flow structure
task	height restriction
expert	rural condition
regulation	turbulent flux of concentration
hazard	furthest upwind surface
framework	inertial sublayer
impact	point of peak influence
guideline	thermodynamic property
delivery	local scale environment
product	explanation of zd
	behaviour of turbulent flux
	microclimate anomaly
	average signal
	local scale surface type
	particular user
	measurement level
	intelligent and flexible way
	frequency of observation
	blending action
	height zr
	extraneous microclimate signal
	temporal scale of interest
	unnatural surface cover
	choice of site
	local surface type
	obstruction of air flow
	ruleofthumb estimate
	nonstandard height
	vertical exchange of momentum
	radiation signal
	homogeneous site
	horizontal effect
	little horizon obstruction
	effect of individual feature
	perspective circle
	short and long wave
	field measurement
	sensor level
	fetch distance
	first surface area

Candidates C1a WMO-No. 1140	Candidates C1b WMO-No. 8
	atmospheric stability
	hygrometer
	surface roughness
	planetary boundary layer
	y direction
	dimension of flux
	downfacing radiometer
	waste heat exhaust
	typical scale of urban microclimate
	climate station recommendation
	standard height of measurement
	meteorological concentration be
	general notion of standardization
	infrared thermometer
	general climatological purpose
	documentation of metadata
	little utility
	nonstandard surface
	cosine response
	single surface cover
	nature of city
	roughness sublayer
	climatic scale
	dimension of individual building
	significant utility
	climate of urban area
	standard climate station
	radiometer signal
	climate observation
	climate of neighbourhood
	respective ellipse
	property of turbulent transport
	action of turbulent eddy
	layer of significant thickness
	rural boundary layer
	atmospheric gas
	specific environment
	automated station
	meteorological environment
	support of detailed forecast
	characteristic mix of microclimatic effect
	urban observation
	meteorological datum
	storm water
	wind engineering

Candidates C1a WMO-No. 1140	Candidates C1b WMO-No. 8
	main roughness element
	microscale effect
	microclimatic effect of individual surface
	sensor location
	mean height
	low density area
	observation of flux
	main property
	meteorological observation
	vertical layer
	surface property
	limited circle
	conceptual illustration
	synoptic space
	climate trend
	microclimate influence
	spacing of building
	urban boundary layer
	similar type of urban development
	surface set
	first order station
	parameter definition
	urban environment
	variable
	distance
	extrapolation
	climate
	signal
	area
	thermometer
	interpretation of datum
	wind direction
	unstable condition
	developed site
	gas analyser
	station site
	essential difference
	stringent standard
	unit of measurement
	usual reason
	anemometer
	ten of kilometre
	minimum distance
	circular patch
	local climate

Candidates C1a WMO-No. 1140	Candidates C1b WMO-No. 8
	stability
	wind
	turbulence
	short distance
	source
	horizon
	concentration
	whole city
	urban design
	site
	humidity
	station
	surface temperature
	obstruction
	hundred of metre
	landscape feature
	property
	vicinity
	radiation
	effect
	heat
	ratio
	exposure
	direction
	flow
	environment
	temperature
	datum
	concept
	obstacle
	location
	air
	requirement
	tree
	type
	building

Annex D – Mathematical calculations for sub-corpora C_{1a} and C_{1b}

	Terms exctracted manually	Terms added to the final termcount after checking out TermoStat results	TOTAL TERMCOUNT on reference list	Candidates Termostat extracted	Terms Termostat extracted	Terms Termostat DID NOT extract	Precision of Termostat		Recall of Termostat	
Corpus C1a - WMO-No. 1140	48	5	53	101	22	31	22/101=0,23	Precision is 23% and Noise is 77%	22/53=0,42	Recall is 42% and Silence is 58%
Corpus C1b - WMO-No. 8	85	1	86	257	59	26	59/257=0,23	Precision is 23% and Noise is 77%	59/86=0,69	Recall is 69% and silence is 31%

Annex E – Reference list for sub-corpora C_{3a} and C_{3b}

Abridged Final Report with Resolutions of the Sixteenth World Meteorological Congress (WMO-No. 1077)	acoustic sounder	sensor
Abridged Final Report with Resolutions of the Sixty-first Session of the Executive Council (WMO-No. 1042)	sodars	sensing element
Abridged Final Report with Resolutions of the Sixty-sixth Session of the Executive Council (WMO-No. 1136)	advection	settlement
aircraft meteorological data relay	aerosol	shelter
AMDAR	air temperature	shielding
aircraft-based observations	air-flow	short-wave radiation flux
barometer	air-stream	long-wave radiation flux
barometer drifter	albedo	sink
BUFR	anemometer	site
Cascading Forecasting Process	anemometer-vane system	siting
CLIMAT	anthropogenic heat	soil temperature
climate monitoring	atmometer	soil moisture
Commission for Basic Systems	atmosphere	solar energy
CBS	atmospheric pressure	solar radiation
Commission for Climatology	atmospheric composition	solar ultraviolet radiation
Consusltative Meetings on Hig-level Policy on Satellite Matters	Atmospheric Radiation Measurement Program	sonic anemometer
Contracting States of ICAO	automatic station	source area
Convention on International Civil Aviation	azimuth	spectrometer
Core Profile	bare ground	static pressure head
CREX	blending height	monitoring station
data buoy	mixing height	Stefan-Boltzmann relation
Data Buoy Cooperation Panel	boom	storm
DBCP	boundary layer	storm front
Data Collection Centres	broadband sensor	streamline
Data Production Centres	narrowband sensor	sunshine
Data Collection Systems	bubble	surface cover
drifter	cavity	surface ice
European Organization for the Exploitation of Meteorological Satellites	urban canopy	surface layer
Executive Council	UCL	surface temperature

Executive Summary of the Abridged Final Report with Resolutions and Recommendations of the Fourth Session of the Joint WMO-IOC Technical Commission for Oceanography and Marine Meteorology (WMO-No. 1093)	urban canyon	surrounding aspect ratio
Expert Team	street canyon	ZH/W
Expert Team on Aircraft-Based Observing Systems	carbon dioxide	synoptic station
Expert Team on Satellite Systems	ceilometer	temperature sensor
extensible markup language	climate station	tethered balloon
XML	clinometer	thunder
geostationary systems	cloud base	tipping-bucket rain gauge
Global Data-Processing and Forecasting System	cloudless sky	tower
Global Framework for Climate Services	collar	transition zones
Global Information System Centre	compass	transmissiometer
Global Observing System	computational fluid dynamics model	turbulence
Global Telecommunication System	debris	turbulence sensor
GRIB	depletion	turbulent flux
Guide for National Meteorological and Hydrological Services Participation in Radio-frequency Coordination	dew	turbulent properties
Guide to Public Weather Services Practices (WMO-No. 834)	dew-point hygrometer	UCZ
Guide to the Global Observing System (WMO-No. 488)	diffuse sky solar radiation	absorption hygrometer
Guide to the WMO Information System (WMO-No. 1061)	direct beam solar radiation	underlying surface
Guidelines for Ensuring User Readiness for New Generation Satellites	dispersion	urban area
Guidelines on data modelling for WMO codes	displacement height	urban boundary layer
ICAO Council	dome	urban hydrologist

Implementation Coordination Team on Information Systems and Services	dust storm	urban site
ICT-IOS	eddy covariance	urban station
Implementation Plan for the Evolution of Global Observing Systems	meteorological element	visibility
EGOS-IP	elevation	von Kármán's constant
integrated observing systems	emissivity	vortex
Inter-Agency Standing Committee	energy balance	wake
Intercommission Coordination Group on the WMO Integrated Global Observing System	evaporation	water budget
Intergovernmental Oceanographic Commission	evaporation pan	water body
IOC	exposure	water flume
International Civil Aviation Organization	extinction coefficient	water cloud
ICAO	fast-response anemometer	weighing gauge
International Meteorological Vocabulary (WMO-No. 182)	fast-response anemometry	water resources
International Telecommunication Union	fetch	wet-bulb sleeve
Inter-Programme Expert Team on Satellite Utilization and Products	fog	wind direction
Joint WMO/IOC Commission for Oceanography and Marine Meteorology	fog bank	wind field
logical data model	fog top	wind load
Manual on Codes (WMO-No. 306)	footprint	wind observations
Manual on the Global Data-processing and Forecasting System (WMO-No. 485)	forcing	wind profiler radar
Manual on the Global Observing System (WMO-No. 544)	friction velocity	wind sensor
Manual on the Global Telecommunication System (WMO-No. 386)	full-scale lysimeter	wind instrument
Manual on the WMO Information System (WMO-No. 1060)	funnel cloud	wind speed
Manual on the WMO Integrated Global Observing System	gas analizers	wind speed gradient
Manual on WIGOS	precipitation gauge	wind tunnel

marine meteorological and oceanographic observing system	raingauge	World Climate Programme
METAR	geographic information system	zenith
moored buoy	Global Atmosphere Watch	open country site
National Centre	Global Energy Balance Archive	heat island effect
National Meteorological and Hydrological Services	temperature gradient	wake effect
NMHSs	gravitational acceleration	incoming radiation flux
numerical weather prediction	gustiness	humidity island effect
observation platforms	gust	minisonde
observing systems	hair hygrometer	wind velocity
Observing Systems Capability Analysis and Review Tool	heat flux	shading ring
OSCAR	heat island	precipitation
OPAG on Integrated Observing Systems	hemispheric reflector	precipitation observation
Radio Regulations	humidity	pressure sensor
radio-frequency spectrum	humidity sensor	propeller anemometer
Regional Telecommunication Hubs	Hybrid Plume Dispersion Model	psychrometric method
Resource Mobilization Office	hygrometer	pyranometer
Satellite Data Telecommunication Systems	incoming direct solar beam	shade ring
SATCOM	incoming flux sensor	shade disc
satellite operator	incoming flux	equatorial mount
satellite system	incoming long-wave radiation	pyrgeometer
Satellite User Readiness Navigator	incoming solar radiation	pyrheliometer
SATURN	inertial layer	radar
Secretariat	inertial sublayer	radiant temperature
Secretary-General	infrared hygrometer	radiation
Severe Weather Forecasting Demonstration Project	infrared remote sensing	radiation balance
SWFDP	instrument enclosure	radiation flux
ship time	internal boundary layer	radiation thermometer
SIGMET	land use	radiative source area
space-based system	laser radar	radio-acoustic sounding systems
surface-based system	lidar	radiometer
SPECI	light wind	radiosonde
standard reporting practices	lightning	rainfall rate
Steering Group of the Severe Weather Forecasting Demonstration Project	local scale	recirculation zones

surface observation	Local-scale Urban Meteorological Parameterization Scheme	remote sensing
upper ocean observation	long-wave radiation	Richardson number
Table Driven Code Forms	low cloud	rime
TDCF	lower UCL	roughness
TAF	lysimeter	roughness elements
Task Team on the Provision of Operational Meteorological Assistance to Humanitarian Agencies	mast	roughness length
Technical Commission	measurement systems	roughness sublayer
Technical Regulations (WMO-No. 49)	mechanical cup anemometer	RSL
Traditional Alphanumeric Codes	mesoscale	rural area
Tropical Pacific Observing System	meteorological instrument	sand storm
United Nations	meteorological observations	satellite obsevation
United Nations Economic and Social Commission for Asia and the Pacific	meteorological optical range	scanner
United Nations Educational, Scientific and Cultural Organization	MOR	scatter meter
warning services	micro-advection	scintillation
weather forecasting	microclimate	scintillometer
WIS centre	micro-lysimeter	screen
WMO Information System	microscale	screen level
WIS	microwave radiometers	seepage
WMO Integrated Global Observing System	microwave temperature profiler	semi-arid site
WIGOS	mixing layer	sensible heat
Regional Association	momentum	sensible heat flux
Workshop on the Impact of Various Observing Systems on Numerical Weather Prediction	Monin-Obukhov similarity theory	phase distortion
Workshop to Assist in Sustaining National Meteorological Services - Strengthening WMO Regional and Global Centres	Obukhov length	planetary boundary layer
World Bank	open country areas	plume
World Meteorological Congress	optical rain gauge	World Radiocommunication Conference
World Meteorological Organization	outgoing radiation	World Weather Watch
WMO	Ozone Limiting Method	radio frequency
operational centre	humanitarian agency	meteorological information

civil aviaiton	meteorological data	logical data-modelling
sea-level pressure	hydrological information	disaster risk reduction

Annex F – Extraction results by MuliTrans of sub-corpora C_{3a} and C_{3b}

regional association	scope of the wmo information system be extended to include	built features
present recommendation	severe weather forecasting demonstration project	ground level
commission for basic systems	direct beam solar radiation	aerial photograph
cbs ext	wmo ioc	voogt and oke
urban areas	set out in part	wind direction
technical regulations	internal boundary layer	cases supporting
noting further	enhancement and expansion of aircraft based observations	requirements of standard reporting practices of table driven code forms
urban stations	support structure	wake effects
roof level	wmo programmes	data availability
air temperature	satcom meetings	highly disturbed
urban station	topographic effects	xml representations of information
urban sites	session of the commission	logical data modelling approach
urban area	guide to the global observing system	expansion of swfdp
observing systems	information system	street canyon
described in annex	marine meteorological and oceanographic	fisheye lens photograph
requests the secretary general	roof tops	element height and density
manual on codes	open construction	humanitarian agencies
satellite systems	international telecommunication union	hydrological and climatological personnel
urban effects	meteorological pre processor scheme	decision making
wind measurements	logical data model	volume i global aspects
annex to the present recommendation	heavily developed urban	recommended in part
solar radiation	commission at its extraordinary session	documented in annex
numerical weather prediction	gauges should be located	meteorological optical range
radio frequencies for meteorological and related environmental activities	global observing	manual on the global observing system wmo
urban districts	part of the wigos	wind instruments
grimmond and oke	manual on the wmo integrated global observing system	amdar programme
national meteorological and hydrological services	urban evaporation	table driven code forms tdcf
urban development	sensible heat flux	executive council at its sixty sixth session agreed

manual on the wmo information system wmo	commission for basic systems relevant	international meteorological vocabulary
wind speed	incoming radiation	implementation plan for the evolution of global observing systems
urban terrain	cause splash	information exchange
wind profile	meteorological and related data and products	part i wigos and the manual on wigos
aircraft based observations	construction activity	sea level pressure
data management	urban soil moisture	surface roughness
contained in annex	relevant to urban areas	blending height
radiation fluxes	sixty fourth session	horizontally and vertically
standard exposure	representative of the local scale urban	physical bulk
wind observations	open country sites	satcom forum
manual on the wmo information system	close attention	developed sites
meteorological data	field of view	authorizes the secretary general to make any consequent
xvi world weather watch programme	regulations for reporting traditional observation data in table driven code forms	contracting states of icao
urban climate	traditional alphanumeric codes	wmo observing systems
radio frequency	global information system centres	joint wmo ioc commission for oceanography and marine meteorology
wind field	missing character strings	careful attention
wmo information system	abridged final report with resolutions of the sixty	net all wave radiation
densely built	temperature sensors	manual on the global data processing and forecasting system
obstacle heights	humidity and wind speed	global and regional
roughness length	decision of the executive council taken at its sixty	speed and direction
meteorological information	surrounding terrain	wmo integrated global observing system wigos
final report	highly variable wind	sensor height
reference height	fifteenth session	hard surfaces
seventeenth congress	development and maintenance	sensing element
open country	source areas	britter and hanna
global telecommunication system	editorial changes	cascading forecasting process
tall buildings	roofs are designed	spatial and temporal

actual evaporation	preserving the radio frequency spectrum for meteorological and related environmental activities at the world radiocommunication conference	radiometer at a height
urban district	satellite data telecommunication systems	satellite data access and exchange
seventeenth world meteorological congress	atmospheric stability	global observing system wmono
climate monitoring	fully documented see section	procedure for adoption of amendments
tropical moored buoy array	global framework for climate services	representative sample
soil moisture	measurement of precipitation	facing radiometers
sketch map	highly variable	low buildings
contained in the manual	radio based	low cloud
flux measurements	sensors in the ucl	demonstration process
urban environment	high level	direction measurements
executive council at its sixty sixth session	tall structures	national meteorological and hydrological services nmhss
short grass	radiative source area	manual on the global telecommunication system wmo
barometer drifter	open water	guide to public weather services practices
mixing layer	incoming long wave radiation	facing radiometer
satellite operators	surface types	amendments to the technical regulations
wave radiation	expert team on satellite	playing fields
standard height	become very hot by day	regional associations and members
fast response	development of the wmo information system	energy balance
urban atmosphere	visibility and mor	session requested the commission
applicable to urban areas	guide for national meteorological and hydrological services participation in radio frequency coordination	boundary layer
short term	effective implementation	scale urban meteorological parameterization scheme
regional associations	hemispheric reflector	urban canopy
wind tunnel	high priority	line represents
integrated observing systems	authorizes the secretary general to make any consequent purely editorial amendments	selected ucz

sixteenth world meteorological congress	hybrid plume dispersion	regional and global
wind sensors	non urban	implementation plans
add an appendix containing	amendments to volume	tall building
international civil aviation organization	plane of the sensing element	designation of centres of the wmo information system
inertial sublayer	horizon obstruction	mixing height
technical regulations wmo	financial burden	underlying surface
united nations	requests the secretary general to arrange for the inclusion of these amendments	data buoy cooperation panel
pollutant dispersion	necessary to describe	based systems
wmo 2003a	bufr or crex	source area
authorizes the secretary general	information in support	low density
trees and buildings	high building	immediate surroundings
guide to the wmo information system	exhaust vents	methods outlined
manual on codes wmo	inertial layer	mechanism to support operational centres
data representations	sponsoring organizations	world weather watch
incoming solar radiation	representative values	radio services
outlined in part	wind profile parameters	global data
urban boundary layer	editorial amendments to the manual	rain gauges
buildings and trees	mast location	abridged final report with resolutions
mean height	recommends the following modifications	davenport and others
consideration by the seventeenth world meteorological congress	height variation	systems applications
amendments to the manual	secretary general to make any consequent purely editorial	height is necessary
urban environments	migration to table driven code forms	height of buildings
climate zone	measurements conducted	urban site
roughness elements	aerial oblique	analysis requires
times their height	key points	displacement height
observing system	remove the conditional reference to centres designations	apply to urban
international civil aviation	ocean observations	microscale sketch
icao council	executive council at its sixty first session	support of international civil aviation
densely developed	competence of meteorological	ucz types
sonic anemometers	buildings analysis	further development
tropical pacific observing system	observations above the rsl	secretariat support

xvi technical regulations of the world meteorological organization	extrabudgetary contributions	models work
amendments to the manual on codes wmo	wmo regions	dbcp implementation strategy
radiation measurement	heat island effect	pan readings
xml representations	requests the secretarygeneral	satellite data collection systems
building height	wmo regional	operational centres
wind speed and direction	meteorological and environmental	air temperatures
technical commissions	remote sensing	united states of america
screen level	sensors should be mounted	systems and services
water pipes	sensor assembly	wis competencies and wis training and learning guide
manual on the global	daytime heat	station metadata
space based	traditional alphanumeric	high rise
surface properties	developed urban area	amendments to icao annex
technical commission	system wmono	flow distortion
instrument enclosure	wmo information system wis	information in extensible markup language
atmospheric properties	seventeenth financial period	regional swfdp projects
expert team	forced ventilation	densely built up area
purely editorial amendments	roughness sublayer	air temperature and humidity
environmental data	manuals and guides	technical report
rain or snow	urban heat	commission for basic systems guidelines for ensuring user readiness for new generation satellites
precipitation observations	open site	radio regulations
wmo recommendations	surface temperature	manual on the wmo information system wmono
height scale	steering group of the severe weather forecasting demonstration project	dust and sand
humidity and wind	operational information	zero plane displacement height
towers and masts	session of the executive council wmo	buildings at relative
open locations	heterogeneity of urban	wmo integrated global observing system
contained in the annex to the present recommendation	cloud base	weather and climate
regulatory material	upper air data coverage	data processing
exchange of meteorological	implementation coordination team	distinct patterns
provision of operational meteorological assistance to humanitarian agencies	requests the secretary general to make the amendments	

Annex G – Extraction results by SynchroTerm of sub-corpora C_{3a} and C_{3b}

abridged final report with resolutions	steering group of the severe weather forecasting demonstration project	radio regulations
addition	streamlines	radio-frequency spectrum for meteorological and related environmental activities at the world radiocommunication conference
advantage	street	radiometer
aerial photograph	street canyon	range
agencies	structure	rate
air temperature	study	readings
aircraft	subject	receipt
albedo	success	recommendations
amdar programme	sufficient	recommendations given
amendments	sun	recommended
amendments to the manual	support	record
amendments to the technical regulations	support structure	reduction
analysis	surface	reference
anemometer	surfaces	reflection
annex	surrogate	region
annex to the present recommendation	surrounding terrain	regional associations
application	surroundings	regional swfdp projects
area	sustainability	regions
array	system	regulations
assessment	table driven code forms	regulations for reporting traditional observation data in table driven code forms
assessment of urban effects	tall building	release
atmosphere	tall structures	report
attention	target	reports
availability	task	representative sample
axis	technical commissions	request
back	technical regulations	requirement
barometer drifter	template	requirements
base	templates	requirements of standard reporting practices of table driven code forms
basis	text	result
benefits	times	results
blending	top	review
booms	tower	risks

building	traditional alphanumeric codes	road
built-up area	transition	roadmap
canopy	tropical moored buoy array	role
capabilities	tropical pacific observing system	roof
cascading forecasting process	types	roof level
case	ucl	roofs
centre	ucz	roughness elements
certain	underlying surface	roughness length
change	united nations	roughness sublayer
characteristics	united states of america	rsl
choice	up	safety
circle	urban area	satcom forum
city	urban atmosphere	scale
climate	urban boundary layer	scientific
climate zone	urban canopy	scope
cloud base	urban climate	screen
collaboration	urban district	secretariat
collar	urban environment	secretariat support
commission	urban site	secretary-general
commission at its extraordinary session	urban station	secretary-general to make any consequent purely editorial amendments
commission for basic systems	urban terrain	selected ucz
commission for basic systems guidelines for ensuring user readiness for new generation satellites	use	sensible heat flux
commission for basic systems relevant	users	sensing element
common	vehicle	sensor
complexity	vents	sensor assembly
complications	vicinity	sensors
component	view	sessions
conditions	wake	set
considerable	way	settlement
construction	wind	seventeenth financial period
context	wind field	seventeenth world meteorological congress
coordination	wind profile parameters	severe
core	wis competencies	severe weather forecasting demonstration project
co-sponsoring organizations	wis competencies and wis training and learning guide	signal
cost	wmo	site
countries	wmo information	siting

data	wmo information system	sixteenth world meteorological congress
data availability	wmo integrated	sixth session
data buoy cooperation panel	wmo integrated global observing system	sketch map
dbcp implementation strategy	wmo programmes	sky
decision	wmo recommendations	soil moisture
decision of the executive council taken	work	source
definitions	workshop	source area
dependence	world meteorological organization	sources
depth	world radiocommunication conference	space
description	world weather watch	spatial
details	world weather watch programme	spatial and temporal
development	xml representations	stable
development and maintenance	year	standard
diameter	years	station
difficulties	zenith	horizontally
difficulty	zero-plane displacement	horizontally and vertically
dimensional	zone	humanitarian agencies
dimensions	measures	icao council
dispersion	metadata	immediate surroundings
displacement	meteorological pre-processor scheme	impact
distance	method	implementation
district	methods	importance
documentation	methods outlined	important role
drop	microscale	inclusion
effectiveness	microscale sketch	increase
effects	minimum	inertial layer
elements	mixing	inertial sublayer
emphasis	mixing layer	influence
enclosure	model	information
end	monitoring	infrared
enhancement	most	instrument
enhancement and expansion of aircraft-based observations	national	instrument enclosure
environment	national meteorological	instruments
environments	national meteorological and hydrological services	interior
error	nature	internal boundary layer
establishment	need	international civil aviation organization

evaporation	needs	international meteorological vocabulary
example	net	international telecommunication union
exchange	network	joint wmo/ioc commission for oceanography and marine meteorology
executive council	new	landscape
executive council at its sixty-sixth session	new systems	large
executive council at its sixty-sixth session agreed	next	last
expansion	nmhss	level
expansion of swfdp	objective	local environment
experience	observation	local scale
expert team	observations	location
exposure	observing systems	logical data model
extraordinary session	obstacles	logical data-modelling approach
fabric	obstruction	loss
fact	obstructions	low density
field	open site	management
field of view	opportunity	manner
fields	organizational	manual
fifteenth session	organizations	manual on codes
final report	own	manual on the global data-processing and forecasting system
financial	pacific	manual on the global observing system
financial burden	pan	manual on the global telecommunication system
first session	pan readings	manual on the wmo information system
fisheye lens	patch	manual on the wmo integrated global observing system
fisheye lens photograph	photograph	manual on wigos
flow	physical bulk	manuals
flux	place	manuals and guides
fluxes	plane of the sensing element	map
following	planning	mast
form	position	mast location
forum	preparation	material
fourth session	pre-processor scheme	mean
fractions	presence	mean wind profile
frequency	present recommendation	measurement
full	procedure	measurement of precipitation

gauge	procedure for adoption of amendments	guide to the global observing system
global	profile	guide to the wmo information system
global data-processing and forecasting system	protection	guidelines
global framework for climate services	provision	gustiness
global observing system	provision of operational meteorological assistance	heat island effect
global telecommunication system	purely editorial amendments	heights
ground	radiative source area	hemispheric reflector
guide	horizon	high
guide for national meteorological and hydrological services participation in radio-frequency coordination	homogeneity	high priority
guide to public weather services practices	historical	

Annex H – Sub-corpora C_{1a} and C_{1b}

These sub-corpora are included for the readers to get an idea of what the entire corpus looks like. They are included only in English.

Sub-corpus C_{1a}:

RESOLUTIONS ADOPTED BY THE SESSION

Resolution 1 (CBS-Ext.(2014))

WMO GUIDELINES ON MULTI-HAZARD IMPACT-BASED
FORECAST AND WARNING SERVICES

THE COMMISSION FOR BASIC SYSTEMS,

Noting that, despite the advances in science and technology, severe weather and associated events still cause many deaths and lead to the destruction of properties and the loss of livelihoods,

Noting further:

(1) That understanding disaster risk and forecasting hydrometeorological impacts is generally beyond the remit of meteorologists and hydrologists and requires close collaboration between National Meteorological and Hydrological Services and partner agencies within government and elsewhere,

(2) That the risk associated with a hydrometeorological hazard depends on knowing how that hazard has an impact on human beings, their livelihoods and assets, which in turn will depend on their vulnerability and exposure,

(3) That the Executive Council discussed the move of National Meteorological and Hydrological Services towards impact-based forecasting and risk-based warnings in the provision of public weather and warning services, in support of social resilience,

Considering the development by the Open Programme Area Group on Public Weather Services of a set of WMO Guidelines on Multi-hazard Impact-based Forecast and Warning Services, as contained in the annex to the present resolution,

Decides to endorse these Guidelines and to request their publication and wide distribution as a means of providing advice and assistance to Members in their move towards impact-based forecasting;

Requests the Open Programme Area Group on Public Weather Services, with assistance from the Secretariat and the regional associations, to provide support to Members in the application of the principles and methodologies described in the Guidelines and on how best to collaborate with their partners in establishing the basis for impact-based forecasting and risk-based warnings.

Resolution 2 (CBS-Ext.(2014))

ESTABLISHMENT OF AN INTERCOMMISSION TASK TEAM TO REVIEW PROCESSES FOR PRIORITIZING OF DATA STREAMS AND CACHE CONTENT

THE COMMISSION FOR BASIC SYSTEMS,

Noting:

(1) Resolution 1 (Cg-XVI) – World Weather Watch Programme for 2012–2015,

(2) The *Manual on the WMO Information System* (WMO-No 1060),

Noting further:

(1) That the capacity of the WMO Information System (WIS) Core Network is a limited resource,

(2) The requirement for Global Information System Centres (GISCs) to ensure the effective operation of the communication systems and services,

Recognizing:

(1) The increasing requirement for the global exchange of all types of environmental data in addition to the established ongoing exchange of meteorological data and products under the auspices of the World Weather Watch and other WMO Programmes, including the Global Framework for Climate Services,

(2) That the addition of new data or products to the WIS Core Network has an impact on all GISCs that have to ensure their capacity is sufficient to handle data and product exchange,

(3) That GISCs have already established a multilateral collaboration forum through the Expert Team on WMO Information System Centres/Task Team on GISCs,

(4) That, although GISCs are responsible for their access to and capacity management of the WIS Core Network, the WMO constituent bodies are ultimately responsible for the content of and capacity requirements for the WMO Information System,

(5) That information that is required to be exchanged between all GISCs is flagged in the discovery metadata by the keyword GlobalExchange,

Decides that there is a need for a formal mechanism for deciding what should be designated for global exchange, and thus held in the cache, and for determining the transmission priority;

Requests the Implementation Coordination Team on Information Systems and Services to establish an ad hoc intercommission task team to work with GISCs to recommend a process for making decisions affecting the capacity management of WIS networks, in particular the WIS Core Network and the GISC cache; the task team should also consider the need for this decision-making process to resolve issues that the collaborative mechanisms established by GISCs are unable to resolve;

Requests the Secretary-General to make available the necessary resources to support the work of this ad hoc intercommission task team.

Resolution 3 (CBS-Ext.(2014))

REGULATION OF CLIMATE DATA MANAGEMENT SYSTEM SPECIFICATIONS THROUGH THE WMO INFORMATION SYSTEM

THE COMMISSION FOR BASIC SYSTEMS,

Noting that the WMO Information System (WIS) provides an overarching approach to data and information management for all WMO and related international programmes,

Noting further:

(1) That the *Manual on the WMO Information System* (WMO-No. 1060), paragraph 3.1, states: "An ongoing process for understanding user requirements … All supported programmes and technical commissions shall participate in this process, which shall be part of general WMO requirement reviews",

(2) That the functional scope and physical sizing of WIS shall be determined through an ongoing process for understanding user requirements,

(3) That the Manual and the *Guide to the WMO Information System* (WMO-No. 1061) are designed to ensure adequate uniformity and standardization in the data, information and communication practices, procedures and specifications employed among Members in the operation of WIS,

Considering:

(1) The recommendation made by the Commission for Climatology (CCl) at its sixteenth session that the Commission for Basic Systems work closely with it in order to identify the relevant parts of the CCl Climate Data Management System (CDMS) Specifications publication for possible inclusion in the WMO Technical Regulations,

(2) That the Manual is Annex VII to the WMO Technical Regulations,

Decides to collaborate with the Commission for Climatology through ad hoc expert meetings to make a proposal for regulating CDMS specifications through WIS regulation and references, including the *Manual on the WMO Information System* and the *Guide to the WMO Information System*;

Requests the Secretary-General to provide support to the collaboration of the Commission for Basic Systems and the Commission for Climatology on the regulation of CDMS specifications.

RECOMMENDATIONS ADOPTED BY THE SESSION

Recommendation 1 (CBS-Ext.(2014))

COMPETENCY FRAMEWORK FOR PUBLIC WEATHER SERVICES FORECASTERS AND ADVISERS

THE COMMISSION FOR BASIC SYSTEMS,

Noting the recommendation of the Sixteenth World Meteorological Congress that all technical commissions make the definition of competence requirements for the core job-tasks in meteorology and hydrology a high-priority activity and incorporate this task into their current work programmes,

Noting further:

(1) That Sixteenth Congress requested that the technical commissions follow the model developed by the Commission for Aeronautical Meteorology in providing top-level competence requirements,

(2) That the Commission for Basic Systems at its fifteenth session decided to establish a Competency Framework for Public Weather Services Forecasters and Advisers,

(3) That this Competency Framework will inform the education and training of forecasters,

(4) That this Competency Framework was developed in close consultation with the Executive Council Panel of Experts on Education and Training and will inform the content of WMO education and training courses, workshops and seminars,

Considering the development by the Open Programme Area Group on Public Weather Services of a Competency Framework that describes the basic forecaster competencies, and then adds on those skills and knowledge that are required for effective service delivery and liaison with users, as contained in sections 1 to 5 of the annex to the present recommendation,

Recommends that this Competency Framework, as contained in the annex to the present recommendation, be endorsed, that it be established as a WMO Recommended Practice, and that it be included in the *Guide to Public Weather Services Practices* (WMO-No. 834);

Requests the Executive Council Panel of Experts on Education and Training to cooperate with the Commission and, with assistance from the Secretariat, to prepare a proposal for amendment of the *Technical Regulations* (WMO-No. 49) to incorporate these competencies as Recommended Practice for the Public Weather Services forecaster/adviser;

Urges Members to make use of this Competency Framework in maintaining and improving the skills and knowledge of staff engaged in the production and delivery of Public Weather Services products and services.

Recommendation 2 (CBS-Ext.(2014))

AMENDMENTS TO THE *MANUAL ON THE GLOBAL DATA-PROCESSING AND FORECASTING SYSTEM* (WMO-No. 485)

THE COMMISSION FOR BASIC SYSTEMS,

Noting the *Manual on the Global Data-processing and Forecasting System* (WMO-No. 485),

Noting further:

(1) The *Abridged Final Report with Resolutions of the Sixteenth World Meteorological Congress* (WMO-No. 1077),

(2) The *Abridged Final Report with Resolutions of the Sixty-sixth Session of the WMO Executive Council* (WMO-No. 1136)

Sub-corpus C_{1b}:

CHAPTER 11

URBAN OBSERVATIONS

11.1 GENERAL

There is a growing need for meteorological observations conducted in urban areas. Urban populations continue to expand, and Meteorological Services are increasingly required to supply meteorological data in support of detailed forecasts for citizens, building and urban design, energy conservation, transportation and communications, air quality and health, storm water and wind engineering, and insurance and emergency measures. At the same time, Meteorological Services have difficulty in making urban observations that are not severely compromised. This is because most developed sites make it impossible to conform to the standard guidelines for site selection and instrument exposure given in Part I of this Guide owing to obstruction of air-flow and radiation exchange by buildings and trees, unnatural surface cover and waste heat and water vapour from human activities.

This chapter provides information to enable the selection of sites, the installation of a meteorological station and the interpretation of data from an urban area. In particular, it deals with the case of what is commonly called a “standard” climate station. Despite the complexity and inhomogeneity of urban environments, useful and repeatable observations can be obtained. Every site presents a unique challenge. To ensure that meaningful observations are obtained requires careful attention to certain principles and concepts that are virtually unique to urban areas. It also requires the person establishing and running the station to apply those principles and concepts in an intelligent and flexible way that is sensitive to the realities of the specific environment involved. Rigid “rules” have little utility. The need for flexibility runs slightly counter to the general notion of standardization that is promoted as WMO observing practice. In urban areas, it is sometimes necessary to accept exposure over non-standard surfaces at non-standard heights, to split observations between two or more locations, or to be closer than usual to buildings or waste heat exhausts.

The units of measurement and the instruments used in urban areas are the same as those for other environments. Therefore, only those aspects that are unique to urban areas, or that are made difficult to handle because of the nature of cities, such as the choice of site, instrument exposure and the documentation of metadata, are covered in this chapter.

The timing and frequency of observations and the coding of reports should follow appropriate standards (WMO, 1983; 1988; 1995; 2003*b*; 2006).

With regard to automated stations and the requirements for message coding and transmission, quality control, maintenance (noting any special demands of the urban environment) and calibration, the recommendations of Part II, Chapter 1, should be followed.

11.1.1 Definitions and concepts

11.1.1.1 Station rationale

The clarity of the reason for establishing an urban station is essential to its success. Two of the most usual reasons are the wish to represent the meteorological environment at a place for general climatological purposes and the wish to provide data in support of the needs of a particular user. In both cases, the spatial and temporal scales of interest must be defined, and, as outlined below, the siting of the station and the exposure of the instruments in each case may have to be very different.

11.1.1.2 Horizontal scales

There is no more important an input to the success of an urban station than an appreciation of the concept of scale. There are three scales of interest (Oke, 1984; Figure 11.1):

(a) Microscale: Every surface and object has its own microclimate on it and in its immediate vicinity. Surface and air temperatures may vary by several degrees in very short distances, even millimetres, and air-flow can be greatly perturbed by even small objects. Typical scales of urban microclimates relate to the dimensions of individual buildings, trees, roads, streets, courtyards, gardens, and so forth. Typical scales extend from less than 1 m to hundreds of metres. The formulation of the guidelines in Part I of this Guide specifically aims to avoid microclimatic effects. The climate station recommendations are designed to standardize all sites, as far as practical. This explains the use of a standard height of measurement, a single surface cover, minimum distances to obstacles and little horizon obstruction. The aim is to achieve climate observations that are free of extraneous microclimate signals and hence characterize local climates. With even more stringent standards, first order stations may be able to represent conditions at synoptic space and time scales. The data may be used to assess climate trends at even larger scales. Unless the objectives are very specialized, urban stations should also avoid microclimate influences; however, this is hard to achieve;

(b) Local scale: This is the scale that standard climate stations are designed to monitor. It includes landscape features such as topography, but excludes microscale effects. In urban areas this translates to mean the climate of neighbourhoods with similar types of urban development (surface cover, size and spacing of buildings, activity). The signal is the integration of a characteristic mix of microclimatic effects arising from the source area in the vicinity of the site. The source area is the portion of the surface upstream that contributes the main properties of the flux or meteorological concentration being measured (Schmid, 2002). Typical scales are one to several kilometres;

Figure 11.1. Schematic of climatic scales and vertical layers found in urban areas: planetary boundary layer (PBL), urban boundary layer (UBL), urban canopy layer (UCL), rural boundary layer (RBL) (modified from Oke, 1997).

(c) Mesoscale: A city influences weather and climate at the scale of the whole city, typically tens of kilometres in extent. A single station is not able to represent this scale.

11.1.1.3 Vertical scales

An essential difference between the climate of urban areas and that of rural or airport locations is that in cities the vertical exchanges of momentum, heat and moisture do not occur at a (nearly) plane surface, but in a layer of significant thickness, called the urban canopy layer (UCL) (Figure 11.1). The height of the UCL is approximately equivalent to that of the mean height of the main roughness elements (buildings and trees), z_H (see Figure 11.4 for parameter definitions). The microclimatic effects of individual surfaces and obstacles persist for a short distance away from their source and are then mixed and muted by the action of turbulent eddies. The distance required before the effect is obliterated depends on the magnitude of the effect, wind speed and stability (namely, stable, neutral or unstable). This blending occurs both in the horizontal and the vertical. As noted, horizontal effects may persist up to a few hundred metres. In the vertical, the effects of individual features are discernible in the roughness sublayer (RSL), which extends from ground level to the blending height z_r, where the blending action is complete. Rule-of-thumb estimates and field measurements indicate that z_r can be as low as 1.5 z_H at densely built (closely spaced) and homogeneous sites, but greater than 4 z_H in low density areas (Grimmond and Oke, 1999; Rotach, 1999; Christen, 2003). An instrument placed below z_r may register microclimate anomalies, but, above that, it “sees” a blended, spatially averaged signal that is representative of the local scale.

There is another height restriction to consider. This arises because each local scale surface type generates an internal boundary layer, in which the flow structure and thermodynamic properties are adapted to that surface type. The height of the layer grows with increasing fetch (the distance upwind to the edge where the transition to a distinctly different surface type occurs). The rate at which the internal boundary layer grows with fetch distance depends on the roughness and stability. In rural conditions, the height to fetch ratios might vary from as small as 1:10 in unstable conditions to as large as 1:500 in stable cases, and the ratio decreases as the roughness increases (Garratt, 1992; Wieringa, 1993). Urban areas tend towards neutral stability owing to the enhanced thermal and mechanical turbulence associated with the heat island and their large roughness. Therefore, a height to fetch ratio of about 1:100 is considered typical. The internal boundary layer height is taken above the displacement height z_d, which is the reference level for flow above the blending height. (For an explanation of z_d, see Figure 11.4 and Note 2 in Table 11.2.)

For example, take a hypothetical densely built district with z_H of 10 m. This means that z_r is at least 15 m. If this height is chosen to be the measurement level, the fetch requirement over similar urban terrain is likely to be at least 0.8 km, since fetch = 100 ($z_r - z_d$), and z_d will be about 7 m. This can be a significant site restriction because the implication is that, if the urban terrain is not similar out to at least this distance around the station site, observations will not be representative of the local surface type. At less densely developed sites, where heat island and roughness effects are less, the fetch requirements are likely to be greater.

At heights above the blending height, but within the local internal boundary layer, measurements are within an inertial sublayer (Figure 11.1), where standard boundary layer theory applies. Such theory governs the form of the mean vertical profiles of meteorological variables (including air temperature, humidity and wind speed) and the behaviour of turbulent fluxes, spectra and statistics. This provides a basis for:

(a) The calculation of the source area (or “footprint”, see below) from which the turbulent flux or the concentration of a meteorological variable originates; hence, this defines the distance upstream for the minimum acceptable fetch;

(b) The extrapolation of a given flux or property through the inertial layer and also downwards into the RSL (and, although it is less reliable, into the UCL). In the inertial layer, fluxes are constant with height and the mean value of meteorological properties are invariant horizontally. Hence, observations of fluxes and standard variables possess significant utility and are able to characterize the underlying local scale environment. Extrapolation into the RSL is less prescribed.

11.1.1.4 Source areas (“footprints”)

A sensor placed above a surface “sees” only a portion of its surroundings. This is called the “source area” of the instrument which depends on its height and the characteristics of the process transporting the surface property to the sensor. For upwelling radiation signals (short- and long-wave radiation and surface temperature viewed by an infrared thermometer) the field of view of the instrument and the geometry of the underlying surface set what is seen. By analogy, sensors such as thermometers, hygrometers, gas analysers and anemometers “see” properties such as temperature, humidity, atmospheric gases and wind speed and direction which are carried from the surface to the sensor by turbulent transport. A conceptual illustration of these source areas is given in Figure 11.2.

Figure 11.2. Conceptual representation of source areas contributing to sensors for radiation and turbulent fluxes of concentrations. If the sensor is a radiometer 50 or 90 per cent of the flux originates from the area inside the perspective circle. If the sensor is responding to a property of turbulent transport, 50 or 90 per cent of the signal comes from the area inside the respective ellipses. These are dynamic in the sense they are oriented into the wind and hence move with wind direction and stability.

The source area of a downfacing radiometer with its sensing element parallel to the ground is a circular patch with the instrument at its centre (Figure 11.2). The radius (r) of the circular source area contributing to the radiometer signal at height (z_1) is given in Schmid and others (1991):

$$r = z_1 \left(\frac{1}{F} - 1 \right)^{-0.5} \quad (11.1)$$

where F is the view factor, namely the proportion of the measured flux at the sensor for which that area is responsible. Depending on its field of view, a radiometer may see only a limited circle, or it may extend to the horizon. In the latter case, the instrument usually has a cosine response, so that towards the horizon it becomes increasingly difficult to define the actual source area seen. Hence, the use of the view factor which defines the area contributing a set proportion (often selected as 50, 90, 95, 99 or 99.5 per cent) of the instrument’s signal.

The source area of a sensor that derives its signal via turbulent transport is not symmetrically distributed around the sensor location. It is elliptical in shape and is aligned in the upwind direction from the tower (Figure 11.2). If there is a wind, the effect of the surface area at the base of the mast is effectively zero, because turbulence cannot transport the influence up to the sensor level. At some distance in the upwind direction the source starts to affect the sensor; this effect rises to a peak, thereafter decaying at greater distances (for the shape in both the *x* and *y* directions see Kljun, Rotach and Schmid, 2002; Schmid, 2002). The distance upwind to the first surface area contributing to the signal, to the point of peak influence, to the furthest upwind surface influencing the measurement, and the area of the so-called "footprint" vary considerably over time. They depend on the height of measurement (larger at greater heights), surface roughness, atmospheric stability (increasing from unstable to stable) and whether a turbulent flux or a meteorological concentration is being measured (larger for the concentration) (Kljun, Rotach and Schmid, 2002). Methods to calculate the dimensions of flux and concentration "footprints" are available (Schmid, 2002; Kljun and others, 2004).