PESCaDO

FP7-248594

Personalized Environmental Service Configuration and Delivery Orchestration

SEVENTH FRAMEWORK PROGRAMME

D7.1 Empirical Study of the Environmental Information Material

Due date of deliverable: 30.06.2010
Actual submission date: 30.06.2010

Start date of project: 1st January, 2010 Duration: 36 months

Lead contractor for this deliverable: UPF Submitted, V1

D7.1	Empirical Study of the Environmental Information Material
Project Acronym :	PESCaDO
Contract No :	FP7 - 248594
Due Date :	30.06.2010
Reply To:	Leo Wanner leo.wanner@upf.edu
Actual date of delivery:	30.06.2010

Deliverable Identification Sheet

Project ref. no.	FP7-248594
Project acronym	PESCaDO
Project full title	Personalized Environmental Service Configuration and Delivery Orchestration
Security (distribution level)	PU
Contractual date of delivery	Month 6, 30.06.2010
Actual date of delivery	Month 6, 30.06.2010
Deliverable number	D7.1
Deliverable name	Empirical Study of the Environmental Information Material
Type	Report
Status & version	Submitted, V1
Number of pages	24
WP / Task responsible	UPF
Other contributors	–
Author(s)	Anton Granvik, Simon Mille, Leo Wanner
Internal Reviewer	Emanuele Pianta and Sara Tonelli, FBK
EC Project Officer	Manuel Monteiro
Abstract	The deliverable summarizes the linguistic analysis of the material provided by FMI and HSY, with the goal to identify the major text, discourse, and syntactic constructions as well as the lexis used to communicate environmental information. In accordance with the priorities of the languages in which the information will be generated in PESCaDO (Finnish > English > Swedish), the focus of the study is on Finnish (which is also due to the circumstance that Finnish is more complex from the information generation point of view than English and Swedish), but English and Swedish are also reviewed in due detail. The results of the study will serve as the basis for the development of text generation resources.

Table of Contents

Executive Summary

In order to be able to generate environmental information in terms the user can understand, it is necessary to study the linguistic constructions experts use to communicate this type of information. The present deliverable analyzes corpora provided by HSY, adding evidence from corpora in the web and from the experience of UPF in the generation of environmental discourse with respect to: 1. textual structures, 2. discourse structures, 3. syntactic structures, and 4. lexical and morphological phenomena. The focus is on Finnish since Finnish is more complex than the other two languages treated in PESCaDO, English and Swedish, from the morphological and syntactical point of view and since Finnish has been hardly worked on so far in the field of information generation (in contrast to, e.g., English). But English and Swedish are also treated in appropriate detail.

The text structures used in environmental discourse are the same in all three languages. They are specific to the genre of generated information: report, advice, and notification (or warning), and the comprehensiveness of the content that is to be communicated. An additional aspect to be considered in the context of text structures is the distribution of the content across the different modi (text, graphics and tables) and the arrangement of the modi in a bulletin.

The argumentation line (i.e., the discourse structure) in environmental discourse is equally language-independent. To model the discourse structure in PESCaDO, we use the Rhetorical Structure Theory (RST), which offers for this purpose a finite set of discourse relations (of the type ELABORATION, CAUSE, PURPOSE, etc.). In the environmental discourse, an identified subset of RST-relations comes into play.

It is the syntactic constructions and lexical and morphological phenomena, which make a difference between the languages. To model syntax and lexical combinatorics, we use the dependency-oriented Meaning-Text Theory (MTT) on which UPF's generator is based. Compared to English (and also to a certain extent to Swedish), Finnish is a relatively free order language, which is reflected first of all in its complex case system. As far as the syntactic constructions are concerned, single clause sentences with locational and temporal circumstantials and coordinated clause sentences dominate. In English environmental discourse, subordination and passivization are frequent. In Swedish, complex sentences with relative clauses are common.

In the discourse of all three languages, a certain number of collocations (i.e., idiosyncratic binary word combinations) are encountered; cf., e.g., *heavy ~ rain*, *weather ~ turn bad*, *sun ~ come out*, etc. It is important to capture (and then use during the process of production) collocations since only they give the bulletins generated by the machine the naturalness needed. To model collocations, we use the typology of lexical functions introduced in MTT.

1 Introduction

As any specialized-discourse reports, environmental information bulletins may contain linguistic constructions that are characteristic to them, i.e., that make them authentic (expert-written). In Natural Language Text Generation (NLG), the knowledge how to write a report in a specific domain is called "domain communication knowledge". The goal of this deliverable is precisely to acquire this domain communication knowledge by carrying out an empirical study, mainly on the multilingual (Finnish, Swedish and English) material produced manually by specialists from FMI and HSY. To complement this material, further material from the web and from other sources is used.
The study covers: 1. the text structure, i.e., the distribution of the information across paragraphs, the order of the paragraphs in different topics and the distribution of the material across different modi (text, table, and graphic); 2. the discourse structure, i.e., the rhetorical organization of environmental information bulletins in terms of discourse relations as introduced in the Rhetorical Structure Theory, RST (Mann and Thompson, 1988); 3. the syntactic sentence constructions following the dependency paradigm, and, more precisely, the Meaning-Text Theory (Mel'čuk, 1988); and 4. the (idiosyncratic) lexical and morphological constructions (with the latter being of higher relevance to Finnish than to English or Swedish).
A particularly relevant aspect of lexical constructions for information production are collocations, i.e., language-specific (= idiosyncratic) binary word combinations, and lexico-semantic derivations. In the deliverable, both are catalogued in terms of lexical functions (Mel'čuk, 1996) – a formal means introduced in MTT for their representation.
A cursory assessment of the corpora reveals that the text and discourse structures of the bulletins largely coincide in all three languages – possibly also because the primary language from which then the bulletins in English and Swedish are derived can be assumed to be Finnish.

2 Text Structure

The user-information corpora provided so far by HSY consist mainly of one paragraph text bulletins on air quality. The corpus survey by FBK (cf. D4.1, due M6) includes web-based daily updated (and downloaded by FBK) information that contains text, graphics and tables, but in its majority focuses on one type of phenomena: air quality, meteorological conditions, or (more seldom) on road conditions.
For the sake of a more flexible and exhaustive information delivery, it should be assumed that PESCaDO-produced bulletins may consist of several paragraphs on a number of related yet different topics. The topics of the bulletins may concern air quality (including pollen), meteorological conditions, and road conditions and be of three different genres: report/briefing, advice/suggestion, and notification (e.g., prompted by fulfilled conditions of a "push" service).
Question answering, which constitutes an important aspect of the PESCaDO-service, can be interpreted either as reporting (when the user, e.g., asks about the current AQ conditions in a given location) or as advising (when the user, e.g., asks whether the current AQ conditions favour a specific outdoors activity).
Each of the above three genres has *per se* a different default text structure – even if in certain contextual settings, they may reveal a similar or even an identical instantiation of it.

2.1 Report Text Structure

The default structure of an exhaustive environmental conditions report involving all three global topics covered by PESCaDO is:

<METEOROLOGICAL CONDITIONS>
<AIR QUALITY CONDITIONS>
<ROAD CONDITIONS>

or

<METEOROLOGICAL CONDITIONS>
<ROAD CONDITIONS>
<AIR QUALITY CONDITIONS>

with each of the blocks being realized as a separate paragraph.
The choice between the two structures depends on the density of the discourse ties between the topics (see Section 3 below). Especially in the case of extreme road conditions resulting from adverse meteorological conditions, the road conditions precede the AQ conditions.
Common to all three topics is the default internal structure in case the user solicits information concerning conditions in the present, future (i.e., forecasts) and/or the past (i.e., archives information): 1. PRESENT, 2. FUTURE, 3. PAST (again, with each of the blocks being realized as a separate paragraph).
In the case of meteorological conditions, the general atmospheric conditions (such as the highs and lows) tend to be communicated in a separate paragraph prior to the concrete conditions (containing temperature, wind, cloudiness, etc.). The concrete conditions are, as a rule, presented in one paragraph.
Air quality information is often to be extended by legal warnings, which are, from the NLG viewpoint, "canned text" to be stored in the database, included without any modification when certain conditions concerning specific air pollution substances are fulfilled. As a rule, these warnings are integrated into the running text on the pollutant in question without that a line break is introduced. If the report is supposed to communicate information on a number of different pollutant substances (e.g., on the relevant pollutants in an area), and this information is not linked by discourse relations, for each of the pollutants a separate paragraph is to be foreseen.
Road conditions are usually communicated in one single paragraph – if the presentation mode is text.

2.2 Advice Text Structure

Advices are assumed to be generated as a reaction to a problem or question formulated by the user. They will be, thus, by definition shorter (and their text structure thus simpler). In general, they are assumed to be one paragraph long if they contain only material that is a direct reaction to the solicitation of the user. If they contain additional information deemed appropriate for communication by the system (e.g., in case the user inquired whether the weather conditions will be appropriate to make a hiking tour in a specific location, and the system possesses the knowledge that the weather is favourable, but the AQ is poor), this information is, as a rule, most appropriately communicated in a separate paragraph.

2.3 Notification Text Structure

A notification concerning the occurrence of environmental conditions will usually have a simple text structure of one paragraph, consisting of a few sentences – such that it is also suitable for being sent via email (or even, when appropriate sentence structures are chosen, as an SMS).

2.4 The Use of Modi

The default communication mode is text. Pictograms are used in connection with road conditions in the case a larger area is covered for which various conditions are to be reported. Graphics in the form of curves are in general to be used for the communication of forecasted or archived conditions for more than 3 time points. Tables are suitable in the case of more than 3 time points for more than 2 subject matters (e.g., pollutants). For illustration, textual information may be also accompanied by a table or graphic. It is thus important to capture the conditions for the use of each of the modi from the text generation viewpoint in order to be able to make in the text references to the content of tables and graphics. Studies involving users will be carried out in the course of the project in which more detailed and reliable criteria for the selection of the appropriate mode under given contextual parameters will be determined.

3 Discourse Structure

As mentioned above, for modelling of the discourse structure, the Rhetorical Structure Theory (RST) is used. Therefore, before typical discourse structures of the environmental reports, advices and notifications are discussed, a brief introduction to the instruments of RST is given.

3.1 Overview of the Rhetorical Structure Theory

RST offers a range of relations (often called coherence relations in linguistic literature) to model discourse. Two major types are distinguished: nucleus-satellite relations and multinuclear relations.

Nucleus-Satellite Relations

The most frequent structural discourse pattern is that two spans of text (virtually always adjacent, but exceptions can be found) are related such that one of them has a specific role relative to the other (or that one depends on the other). For instance, a discourse can contain a claim followed by evidence for this claim. RST introduces an “Evidence” relation between the two spans. It also says that the claim is more essential to the text than the particular evidence, and this essentiality is represented by calling the claim span a *nucleus* and the evidence span a *satellite*. About twenty of such nucleus-satellite relations are distinguished.

The order of spans in nucleus-satellite relations is not constrained, but there are more likely and less likely orders for all of the relations. A list of main relations used in RST, and a short description of the possible nuclei and satellites is given in Table 1:

Relation Name	Nucleus	Satellite
Antithesis	ideas favored by the author	ideas disfavored by the author
Background	text whose understanding is being facilitated	text for facilitating understanding
Circumstance	text expressing the events or ideas occurring in the interpretive context	an interpretive context of situation or time
Concession	situation affirmed by author	situation which is apparently inconsistent but also affirmed by author
Condition	action or situation whose occurrence results from the occurrence of the conditioning situation	conditioning situation
Elaboration	basic information	additional information
Enablement	an action	information intended to aid the reader in performing an action
Evaluation	a situation	an evaluative comment about the situation
Evidence	a claim	information intended to increase the reader's belief in the claim
Interpretation	a situation	an interpretation of the situation
Justify	text	information supporting the writer's right to express the text
Motivation	an action	information intended to increase the reader's desire to perform the action
Non-volitional Cause	a situation	another situation which causes that one, but not by anyone's deliberate action
Non-volitional Result	a situation	another situation which is caused by that one, but not by anyone's deliberate action
Otherwise (anti conditional)	action or situation whose occurrence results from the lack of occurrence of the conditioning situation	conditioning situation
Purpose	an intended situation	the intent behind the situation
Restatement	a situation	a reexpression of the situation
Solutionhood	a situation or method supporting full or partial satisfaction of the	a question, request, problem, or other expressed need

	need	
Summary	text	a short summary of that text
Volitional Cause	a situation	another situation which causes that one, by someone's deliberate action
Volitional Result	a situation	another situation which is caused by that one, by someone's deliberate action

Table 1: Most common nucleus-satellite relations in RST

Multinuclear Relations

In addition to the most frequent pattern of nucleus and satellite, there are relations that do not carry a definite selection of one nucleus. These are called *multinuclear relations*, summarized in Table 2:

Relation Name	**Span**	**Other Span**
Contrast	one alternate	the other alternate
Joint	(unconstrained)	(unconstrained)
List	an item	a next item
Sequence	an item	a next item

Table 2: Multinuclear relations in RST

From the perspective of the use of the discourse relations in information generation, it can be stated that in the first step of generation, a document plan has to be mapped to a conceptual representation which serves as the actual input to the text generator. The document plan consists of information spans and rhetorical (or *discursive*) relations which connect the spans. The information spans are provided by the content selection module, while the rhetorical relations are introduced by the document planner.

3.2 Discourse Structures in Environmental Discourse

In the environmental discourse, a subset of discourse relations listed above are commonly used – among them, the nucleus-satellite ELABORATION, JUSTIFY, PURPOSE, and INTERPRETATION and the multinuclear CONTRAST, JOINT, LIST, and SEQUENCE.

An information span in the environmental discourse usually consists of one statement (in linguistic generation rendered into a single sentence). However, it may also contain a configuration of statements – in the same vein as several information spans can be rendered via aggregation into a single sentence.

Figure 1 displays an example of an RST-based discourse structure of the environmental discourse. Thus, between the proposition on the air quality index and the position identifying the primary pollutant a cause relation holds. Statements identifying the secondary pollutants might elaborate the information on air quality further. Between the

statements on the threshold and the pollutants which exceed the threshold an evaluation relation exists.
Between the statement forecasting the concentration of a pollutant substance and the statement introducing this substance, an ELABORATION relation holds. Between the reasons that led to a specific forecast and the forecast itself, the EVIDENCE relation holds.

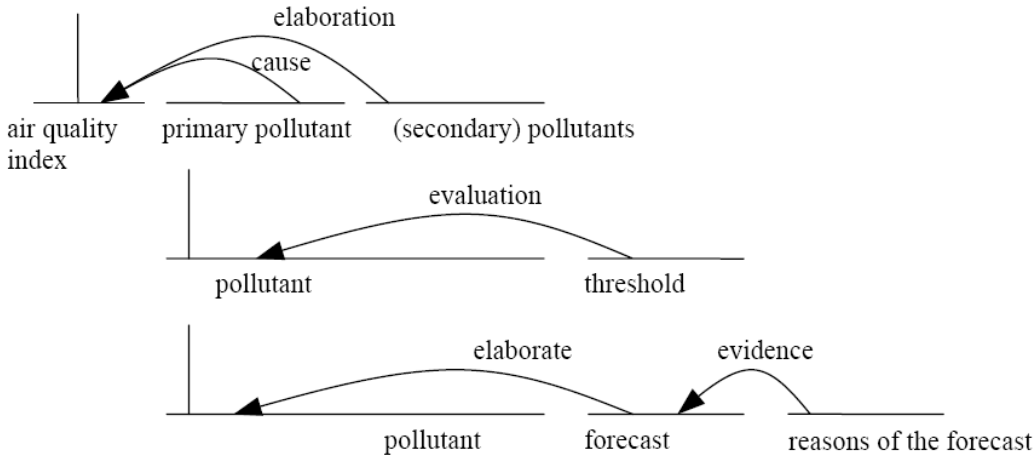

Figure 1: General discourse relation schema

Some information spans (such as, e.g., secondary air pollutants) are less relevant to the user. Nonetheless, they are occasionally included into the discourse – for instance, to contrast them to the primary pollutants and thus achieve a stronger emphasis of the latter; cf. Figure 2.

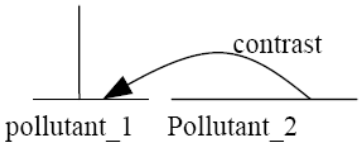

Figure 2: CONTRAST relation between statements on secondary and primary pollutants

When two pollutants have the same relevance to the user, the SEQUENCE relation is used to communicate on them.

Using discourse particles to achieve a coherent text
Some of the discourse relations can be signaled explicitly in the text by specific discourse particles – such as, e.g., JUSTIFY by *therefore*, INTERPRETATION by *this means*, CONCESSION by *however*, etc. The selection of a discourse particle in a specific context may also restrict the order of the sentences and vice versa. Consider some examples:

CAUSE(N, S):
Typical discourse particles: *due to* (*that*), *as*, *the reason for*, *because of*, etc.
Example:
Nucleus: *The PM10 concentration is high.*
Satellite: *The air quality is bad.*

Possible realizations:
1. *Due to the high PM10 concentration, the air quality is bad*.
2. *The air quality is bad. The reason for that is the high PM10 concentration*.
3. *The air quality is bad as the PM10 concentration is high*.
4. *The high PM10 concentration is responsible for the bad air quality*.

CONTRAST(N, N)
Typical discourse particles: *however*, *but*, *while*, etc.
Example:
Nucleus1: The NOx concentration is high.
Nucleus2: The ozone concentration is low.
1. *The NOx concentration is high, while the ozone concentration is low*.

SEQUENCE(N, N, …)
Typical discourse particles: *and*, *or*
Example:
Nucleus1: The PM10 concentration is high.
Nucleus2: The ozone concentration is low.
1. *The PM10 and ozone concentrations are high*.
2. *The weather will be in general sunny and the air quality good*.

4 Syntactic Constructions

4.1 Finnish

4.1.1 Word Order

Word order in Finnish is free, especially in comparison with English or other Germanic languages like Swedish. The role of the noun is determined not by word order or sentence structure but by case markings, which indicate subject, object and indirect object. There is, however, no specific Dative case for the IO. Instead, one of the locative cases is used, namely the allative 'to'-case. The usual, neutral word order is SVO. Subject marking is always realized morphologically as a verb ending, which means that overt subject marking is optional except for the third person. This means that, in general, if the subject is third person, it cannot be omitted, except for cases such as *sataa* 'it rains' and other impersonal expressions (such as the Finnish passive or impersonal construction).
Adverbial phrases or elements can be inserted quite freely, but complements —often realized as complex NPs involving nominalizations in genitive case, much like Spanish *de*-complements— normally precede their head. Relative sentences are used rather seldom in the environmental reports, precisely because of the frequent use of anteposed genitival complements. An example:

Katupölyn	*vähentämiseksi*	*toteutetut*	*toimet*	*vaikuttavat*
'street dust-GEN	reduction-TRA	take-Ppcpl-PL	action-PL	'work-PRES-3.p.PL'

'the actions taken in order to reduce street dust are effective'

4.1.2 Sentence structure and complexity

The general sentence type of the Daily Air Quality Assessments are copulative sentences describing the air quality in the Helsinki metropolitan area. A prototypical sentence looks as follows: *Ilmanlaatu on hyvä suurimmassa osassa pääkaupunkiseutua*

'AQ is good in most parts of the metropolitan area'. Following the example just presented, a large number of sentences are simple (i.e. uncoordinated) main clause sentences consisting of subject, predicative and adverbials (of varying kinds). Another highly frequent clause type is the existential: *Ilmassa on liikenteen pakokaasuja aamuruuhkan aiheuttamana* 'There is traffic exhaustion in the area due to the morning rush hour'.

Complex sentence types are fairly rare in the corpus studied, but at least coordinate sentences occur with a certain frequency, e.g. *Ilmanlaatu on välttävä Helsingin kantakaupungissa ja muualla seudun vilkasliikenteisillä alueilla* 'AQ is acceptable in the Helsinki city centre and in other densely trafficated areas. What is coordinated here are two locative expressions, so the actual sentence is not very complex. Subordinate clauses are rare in the corpus under study, but relative clauses, subordinate adverbials ('because') and *if*-clauses, do occur. Here are some examples of which only the first type occurs repeatedly: *Töölöntullissa, jossa ilmanlaatuindeksi on tyydyttävä* 'In Töölöntulli, where the AQ index is satisfactory', a locative relative; *HSY, joka jatkaa* 'HSY, who continues...'; and, finally, *Saasteiden määrä pysyy ilmassa kohtuullisena, koska tuuli puhdistaa ilmaa* 'The amount of pollution in the air remains at a reasonable level since the wind cleans up the air'.

A highly frequent, recurring syntactic pattern of the Finnish AQ reports is the accumulation of adverbials expressing the circumstances in which the different measurements have been made. These complements can be divided into locative, temporal, causative and what we have described as background adverbials. The temporal, causative and background adverbials are fairly uncomplicated and usually occur only once in each sentence. Here are a couple of illustrating examples: *Ilma on* ***tänä aamuna*** *melko puhdasta* '**This morning** the air is fairly clean' and *Ilmanlaatu on paikoin välttävä* ***pakokaasujen takia*** 'At places air quality is weak **because of exhaustion**'. The locatives are in principle not unlike the above examples but, interestingly enough, they tend to appear in groups making up an interesting hierarchy of spatial inclusion. This marked tendency is further enhanced by the fact that many locatives bear the same case ending, making the structures syntactically ambiguous. The following example is highly illustrative:

> ***Muualla pääkaupunkiseudulla*** *liikenteen pakokaasut heikentävät ilmanlaadun* ***vilkasliikenteisillä alueilla paikoin*** *välttäväksi.*
>
> 'In other parts in the metropolitan area the exhaust gases from the traffic reduce air quality to poor at places in densely trafficated areas.'

In this example, the sentence begins with a double locative, *muualla* 'in other parts' and *pääkaupunkiseudulla* 'in the metropolitan area'. The relation between these two locative elements is obvious, but the syntactic relation is one of coordination rather than inclusion due to the fact that both bear the same adessive case ending (*-lla*). One can also omit either of the two without changing the meaning of the sentence. In a similar fashion, the sentence ends with two more locatives, *paikoin* 'at places' and *vilkasliikenteisillä alueilla* 'in densely trafficated areas'. Here twe two elements bear different locative cases (as reflected in the translation, *at* vs. *in*), the main locative bearing the adessive case, while *paikoin* is to be considered an invariable adverb. The difficulties as to how to determine the relation between them both remain: *paikoin* can be understood as referring to a location included in the 'densely trafficated areas', but this is, of course, a semantic interpretation. Syntactically both elements are independent and either one can be removed without significantly modifying the meaning of the sentence.

4.1.3 Particles and clitics

Since the Finnish derivational system is very rich, verb particles are not very numerous. Some are, however, quite frequently used and appear also in the annotated environmental material. Worth mentioning are, e.g. *pois* 'away', *mukaan* 'with' and *vastaan* 'against' as in *mennä pois* 'to go away', *tulla mukaan* 'to come with' and *tulla vastaan* 'to come against=to meet or to concede'.
Typical for Finnish is also the use of clause or sentence level clitics, such as interrogational *-kO*, additive *-kin* and negative *-kAAn*, as well as verifying *-han* (cf. the examples with the verb *istua* 'to sit' above). The interrogative suffix can be attached to any word class forming both direct and indirect questions and is obligatory expressed on the verb in normal, unmarked finite interrogative sentences which lack an interrogative pronoun: *haluat**ko*** auton? 'Do you want a car?' The suffixes *-kin* and *-kAAn* have functions similar to English 'too' or adverbs like *myös* 'also/neither', Sp. 'también/tampoco'. They are also obligatory on some indefinite pronouns such as *kukin* 'each one', *(ei) kukaan* 'no one', *(ei) mikään* 'nothing'.

4.2 English

The English language is much simpler than Finnish, described in the previous sections. In this section a brief overview of its syntax as found in the corpus is given.
In the English corpus, one clause sentences, usually with some temporal, locative and manner circumstantials dominate. The sentences are usually declarative; there are no questions or exclamations, for instance.
EX: *This morning, the PM10 concentrations were above the information threshold.*

English is a Subject-Verb-Object language in that almost every declarative sentence has this basic structure with this basic order. The subject has to be expressed, and sometimes triggers an agreement on the verb (see next section about morphological issues).
The sentences can be combined using coordination (*and*, *or*, *but*, etc.) or subordination, thanks to subordinating conjunctions, such as *when*, *because*, etc., or relative pronouns (*that*, (in) *which*, etc.).
There is no impersonal voice in English, but the passive constructions -auxiliary *be +past participle-* can be used instead.
Ex: *Perturbations in the traffic in the center of Helsinki **are** expect**ed**.*

The order of words in English is quite fixed, unlike Finnish. However, the word order can be used to emphasize information in a proposition, by fronting the emphasized information to the sentence top (in (b) below).
EX: (a) *The air quality is rather good **this morning**.*
VS
*(b) **This morning**, the air quality is rather good.*

In order to avoid repetitions in the texts, noun groups can be pronominalized thanks to personal pronouns in the position of syntactic subjects (*it*, *they*, etc.) or objects (*it*, *them*, etc.).
EX: *This morning, the air quality is rather bad, whereas yesterday **it** was very good.*

As far as verbal tenses are concerned, in the domain occur only sentences with three different tenses. The present tense is used for propositions that contain recent

concentrations or propositions about information today. The past tense can be found in propositions of past measurement or in propositions about time prods bygone such as *yesterday*, *in the morning*, *in the afternoon*, etc. Future is used in forecasts.

Negation in English is simply expressed by an adverb *not* on the concerned verb, or *no* on a noun.

EX: *Yesterday, the PM10 concentration was **not** measured.*
VS
*There was **no** measurement of the PM10 concentration yesterday.*

4.3 Swedish

4.3.1 Word order

Word order in Swedish is, as in many other Germanic languages and contrary to Finnish, quite restricted, with one essential rule: the main verb of a declarative sentence must always come in second place (V2-language), e.g. *Det svävar alltid partiklar i luften.* 'There float always particles in the air'. However, in subordinate phrases introduced by a subordinate conjunction (*if*, *when*, *since*,…) the finite verb comes in the third place, the subordinator falling, so to speak, outside the basic phrase structure. The basic word order SVO including prepositions and preposed modificators. Questions are formed by inversing the (initial) word order *Lufkvalitén är god idag* vs. *Är luftkvalitén god idag?* 'AQ is good today vs. Is AQ good today?'

4.3.2 Sentence structure and complexity

Swedish generally uses a quite simple sentence structure with restricted use of complex subordinate clauses. Relative clauses are frequent instead of heavy accumulation of adjectives. Nominalizations are not to be preferred but do occur with some frequency in the Finnish meteorological texts, presumably as a reflection of their Finnish origin.

4.3.3 Particles

Swedish is a so-called satellite-framed language, meaning that verbs generally do not encode direction but this and other notions, such as aspect, is expressed by means of so called verb particles that are attached to the verb, e.g. *gå ut* 'go out' vs. Spa. *salir* or *komma tillbaka* 'return; lit. come back', 'volver'. The verb particles often coincide with the most general prepositions, e.g. *titta på teve* 'watch tv' and *på landsbygden* 'in the countryside' where the same element, *på* is used both as verb particle and as a locative preposition.

5 Lexical and Morphological Issues

5.1 Finnish

Finnish is the language that poses most of the challenges in PESCaDO in that it belongs to a family not yet dealt with widely in NLG and in that it reveals considerable complexity as far as its morphology and morpho-syntax is concerned, we discuss its characteristics in some detail.

5.1.1 Vowel harmony

Vowel harmony is a redundancy feature, which means that the feature [±back] is uniform within a word, and so it is necessary to interpret it only once for a given word. It is meaning-distinguishing in the initial syllable, and suffixes follow; so, if the listener hears [±back] in any part of the word, they can derive [±back] for the initial syllable. For example, *tuote* ("product") agglutinates to ***tuo**tteeseens**a*** ("into his product"), where the final vowel becomes the back vowel 'a' (rather than the front vowel 'ä') because the initial syllable contains the back vowels 'uo'. This is especially notable because vowels 'a' and 'ä' are different, (meaning-distinguishing) phonemes, i.e. not interchangeable or allophonic.

5.1.2 Lexis

5.1.2.1 Word formation

Finnish extensively employs regular **agglutination**. It has a smaller core vocabulary than, for example, English, and uses derivative suffixes to a greater extent. Here are some of the more common such suffixes. Which of each pair is used depends on the word being suffixed in accordance with the rules of vowel harmony.

- ja/jä : **agent** (one who does) (e.g. käyttää 'to use' → käyttäjä 'user')
- lainen/läinen: **inhabitant of** (either noun or adjective). Helsinki → helsinkiläinen 'from Helsinki'; Suomi → suomalainen 'Finnish person or thing'.
- sto/stö: **collection of**. For example: mies 'a man' → miehistö 'personnel'; laiva 'a ship' → laivasto 'navy, fleet'.
- ton/tön: **lack of something**, 'un-', '-less' (savu 'smoke' → savuton 'without smoke'; koti 'home' → koditon 'homeless').

Verbal suffixes are extremely diverse; several frequentatives and momentanes differentiating causative, volitional-unpredictable and anticausative are found, often combined with each other, often denoting indirection. For example, *hypätä* 'to jump', *hyppiä* 'to be jumping', *hypeksiä* 'to be jumping wantonly', *hypäyttää* 'to make someone jump once', *hyppyyttää* 'to make someone jump repeatedly' (or 'to boss someone around'), *hyppytyttää* 'to make someone to cause a third person to jump repeatedly', *hyppyytellä* 'to, without aim, make someone jump repeatedly', *hypähtää* 'to jump suddenly' (in anticausative meaning), *hypellä* 'to jump around repeatedly', *hypiskellä* 'to be jumping repeatedly and wantonly', *hyppimättä* 'without jumping', *hyppelemättä* 'without jumping around'. Often the diversity and compactness of this agglutination is illustrated with *istahtaisinkohan* 'I wonder if I should sit down for a while' (from *istua*, 'to sit, to be seated'):

istua 'to sit down'
istun 'I sit down' / *istahtaa* 'to sit down for a while'
istahdan 'I sit down for a while'
istahtaisin 'I would sit down for a while'
istahtaisinko 'should I sit down for a while?'
istahtaisinkohan 'I wonder if I should sit down for a while'.

5.1.2.2 Collocations

A number of typical collocations (i.e., language idiosyncratic binary word combinations) are encountered in the Finnish environmental discourse; cf. the table

below. As mentioned above, collocations will be captured by lexical functions (LFs). When applied to a keyword (= the *base*), an LF provides a set of values (= the *collocates*) that form together with the base collocations with the meaning denoted by the LF itself. In the table below, LFs are presented in the last column ($Oper_1$, $Oper_0$, Magn, AntiMagn, …).

Sentence/predicate types	*Ilmanlaatu on [adv] hyvä/tyydyttävä/välttävä/huono*	AQ is [adv] good/fair/passable/bad	$Oper_1$(hyvä...)
	Ilmassa on pakokaasuja	There is exhaustion in the air.	$Oper_0$(pakokaasuja)
	Sää on tyyni/tuulinen	The weather is calm/windy	$Oper_1$(tyyni)
	Ilma on puhdasta	The air is clean	$Oper_1$
	Tuuli on heikkoa/voimakas/navakka	The wind is soft/heavy/hard	Magn AntiMagn (tuuli)
	Saastepitoisuudet ovat matalia	Pollution densities are low	AntiMagn (saastepitoisuudet)
Modifying structures	*Vilkkaasti liikennöityjen katujen*	Lively trafficated streets Densely trafficated area	Magn(liikennöity)
	Tuulinen sää *Heikkotuulinen sää* *Voimakas tuuli* *Navakka tuuli*	Windy weather Lightly windy weather	AntiMagn(tuulinen)
	Heikentävät ilmanlaadun välttäväksi.	Reduce AQ to fair	Bon(ilmanlaatu)
	Ilmanlaatu vaihtelee tyydyttävästä hyvään	AQ varies from fair to good	Bon(Bon(ilmanlaatu))

Table 3: Common collocations in Finnish environmental discourse

5.1.3 Morpho-Syntax

5.1.3.1 Verbs

The morphosyntactic alignment is nominative-accusative; but there are **two object cases**: accusative and partitive. The contrast between the two is telic: whereas the accusative case denotes actions completed as intended (*Ammuin hirven* 'I shot (killed) the elk'), the partitive case denotes incomplete actions (*Ammuin hirveä* 'I shot (at) the elk'). Transitivity is distinguished by different verb forms (arrived at derivationally) for transitive and intransitive meaning/use, e.g. tr. *ratkaista* 'to solve something' vs.itr. *ratketa* 'to be solved by itself. There are several frequentative and momentane verb categories.

Verbs gain **personal suffixes for each person**; these suffixes are grammatically more important than pronouns, which are often not used at all in standard Finnish. The **infinitive** is not the uninflected form but has a suffix *-ta* or *-da*; the closest one to an uninflected form is the third person singular indicative. There are **four persons**, first ("I, we"), second ("you (singular), you (plural)"), third ("s/he, they"). The **passive** voice (sometimes called impersonal or indefinite) resembles a "fourth person" similar to e.g. English 'people say/do/…'. There are **four tenses**, namely present, past, perfect and pluperfect; the system mirrors the Germanic system.

Finnish does not have a separate verb for possession. Possession is indicated in other ways, mainly by genitives and existential clauses. For animate possessors, the adessive case is used with 'olla', for example *palveluntuottajalla on ongelmia* = 'the service producer has some problems' - literally 'on the service producer is problems'.
Indicating possession is one of the numerous uses of the verb ***olla*** 'to be'. Other functions of *olla* are existencial 'there is/are' expressions, in which the verb invariantly takes the third person singular form, *on*. Besides expressing existence, *olla* also has a locative meaning, used for situating people and things, very much like English be *hän on ulkona* 'he is outside'. Inflected *olla* is further used as the prefect and pluperfect auxiliary: *olen/olin tehnyt* 'I have/had done'.

Negation

Verbs are negated by using a 'negative verb' in front of the verb (in the present, the stem from the present tense in its 'weak' consonant form; in the past tense the "participe" is used): *en, et, ei, emme, ette, eivät mene* 'I not,you not,s/he not, we not, you not, they not go'.

Passive voice

The Finnish passive is unipersonal, that is, it only appears in one form regardless of who is understood to be performing the action. In that respect, it could be described as a "fourth person", since there is no way of connecting the action performed with a particular agent (except for some nonstandard forms). Notice also that the object is usually in the nominative case. However, verbs which govern the partitive case continue to do so in the "passive", and where the object of the action is a personal pronoun, it bears the direct object specific accusative form: *minut* unohdettiin 'I was forgotten'. For the purposes of the environmental reports, the single complement of the unipersonal verb form is considered to be a direct object.

Infinitives

Finnish verbs are described as having four, sometimes five infinitive forms. These are all derived forms of the verb which exhibit mainly nominal uses.

The **first infinitive** short form of a verb is the "dictionary entry" form. It is not unmarked; its overt marking is always the suffix *-a* or *-ä*. This form is the one typically used in complement clauses much like English or Spanish infinitives: *quiero cantar, I want to sing = haluan laulaa*.

The **second infinitive** is used to express aspects of actions relating to the time when an action takes place or the manner in which an action happens. In equivalent English phrases, these time aspects can often be expressed using 'when', 'while' or 'whilst' and the manner aspects using the word 'by' or else the '-ing' form of the English verb to express manner.
It is recognizable by the letter 'e' in place of the usual "a" or "ä" as the infinitive marker. It is only used with one of two case makers: the inessive "-ssA" indicating time, or the instructive "-n" indicating manner. Finnish phrases using the second infinitive can often be rendered in English using the "-ing" verb form.
The second infinitive is formed by replacing the final 'a'/'ä' of the first infinitive with 'e' then adding the appropriate inflectional ending: 'tehd**e**ssä' 'when doing', 'sano**e**ssa' 'when saying'. Adding a possessive suffix to the inflected 2nd infinitive includes the semantic subject: *sanoessani* 'When I am/was saying'.

This corresponds to the English verbal noun (-ing form), and behaves as a noun in Finnish in that it can be inflected, but only in a limited number of cases. It is used to refer to a particular act or occasion of the verb's action.
The **third infinitive** is formed by taking the verb stem with its consonant in the strong form (a), then adding 'ma/mä' followed by the case inflection.
The cases in which the third infinitive can appear are:

tekemässä	inessive	'in doing'
tekemästä	elative	'from doing'
tekemään	illative	'to doing'
tekemällä	adessive	'by doing' (mean)
tekemättä	abessive	'without doing'

The **fourth and fifth infinitives** are rarely used.
Though not an infinitive, a very common *-minen* verbal stem ending results in a noun phrase which gives the name of the activity described by the verb (e.g. *sekoittuminen*, 'mixing' as in *ilman sekoittuminen on heikkoa tyynen sään vuoksi* 'Mixing of the air is little due to calm weather'. This is rather similar to the English verbal noun *-ing* or romance nominalizations in *-tion*/*-ción*. As a noun, the nominalisation in *-minen* inflects just like any other noun, e.g. *ilman sekoittumise**lla** tarkoitetaan...* 'by (using the term) mixing of the air we mean...' (see Section 4.1 for another example).

5.1.3.2 Nouns

Finnish does not distinguish **gender** neither for nouns nor for pronouns and no distinction is made between **determinate and indeterminate forms**, e.g. the NOM noun form *talo* can be translated by both 'a house' and 'the house'. There are **two numbers**, singular and plural, from which only the second bears morphological marking. The plural, however, is marked differently according to context. The nominative plural is a definite, telic plural and is marked by final *–t*, *päästö – päästöt* 'exhaustion – exhaustions'. The inflected plural adds the ending *-i-* before the singular partitive ending, e.g. *päästö-j-ä* 'exhaustions'. Finally, with numerals higher than one, the singular partitive is used, e.g. *kaksi kertaa ≠ useita kertoja* 'two timer ≠ many times'.
Subjects and direct objects nouns may be suffixed with the markers for the nominative (unmarked), and for genitive and partitive case, respectively. Accusative case exists only for personal pronouns. Some verbs require that subjects be marked by the genitive, such as modal *pitää* 'have to'. In negated clauses the default case for marking the direct object is the partitive.
Beside these "argument cases" there are six different **locative cases** (rough equivalents to the notions *in, on, at, by, from, out of, with*), as well as **five other cases** with more limited uses. (For the complete list of names and endings, see http://kaino.kotus.fi/visk/sisallys.php?p=81.) Case marking is added not only to the main noun, but also to its modifiers; e.g. *suure+ssa talo+ssa*, literally "big-in house-in" = 'in the big house.
Possession is marked with a possessive suffix attached to the possessed noun. There are no separate possessive pronouns, but the personal pronouns in the genitive case function as such. Some examples: the construction "my book" can be expressed in two ways: *kirja+ni = kirjani*, or *minun kirja-ni*, where *minun* is GEN of the personal pronoun *minä* 'I'. That is, the Finnish equivalents of the possessive pronouns, e.g. *my, your, her/his, our...* are the genitive forms of the personal pronouns, that is, *minä -> minun, sinä -> sinun, hän -> hänen...* Most pronouns take case suffixes just as nouns do.

An interesting, and challenging, part of Finnish noun morphology is the possibility to create **compounds**. The structure is generally such, that the head is complemented by prefixing the modifying part, e.g. *ilmanlaatu* 'air quality'. The challenge lies in two particular details: Firstly, the complement part can often be attached to the head in an unmarked, nominative case like form, e.g. *raja-arvo* 'limit value'. Quite often, though, the genitive case is marked also on the complementing word, as in *ilma**n**laatu* (*ilma+n*). In some cases, depending on the complementing element, abbreviated forms are used. This is especially the case of adjectives, such as *pientaloalue* 'small house area' where the adjective *pieni* 'small' appears in a shortened form. As this example shows, compounds consisting of multiple elements are possible. Secondly, it is not at all certain where one is to draw the limit between what constitutes a compound proper (analyzable as a composite unit with distinguishable parts) and what constitutes a unit which does not easily lend itself to analytical decomposition. In the above example, *pientaloalue*, the first compound *pientalo* is certainly a particular kind of house, but in the case of *rivitalo* 'row house' or the even longer *omakotitalo* 'detached house/particular house for one family' it is clearly a question of conceptually discrete units.

The environmental domain is rich on compounds and collocations, some of which relate explicitly to the environmental domain, such as *ilmanlaatu* 'air quality'*, pakokaasuja* 'exhaustion'*, katupölyä* 'street dust'*, hiukkaspitoisuudet* 'concentration of particles*, ilmanlaatutietoja* 'air quality information'. Many locative expressions and some frequent adjectives are also compound forms, e.g. *pääkaupunkiseutu* 'the metropolitan area', *pääväylät* 'main roads', *ydinkeskusta* 'the city centre', *hengitysilma* 'inhaled air', etc. Among the adjectives v*ilkasliikenteinen* 'densely trafficated', *heikkotuulisen* 'lightly windy' and *hienojakoista* 'fine(ly divided)' can be mentioned.

5.1.3.3 Adjectives

Adjectives take the same case endings as their head noun. This includes the comparative and superlative forms, which first take the corresponding ending, *-mpi* and *-in* respectively and then case suffixes are added.

5.1.3.4 Postpositions and prepositions

Postpositions are more common in Finnish than prepositions. Both postpositions and prepositions can be combined with a noun to form an "adpositional" phrase, e.g. *talon takana* "behind the house". However, with personal pronouns two alternatives exist, the normal way, by using the genitive form of the personal pronoun, *minun takana* 'behind me' or by adding the possessive suffix to the adposition, giving *takana**ni*** 'behind me'. Postpositions indicate place, time, cause, consequence or relation. In postpositional phrases, the noun is usually in genitive. Many postpositions are actually inflected nouns –semantically transparent– whose syntactic function has become fixed introducing adverbial expressions. In this sense, the adpositions can be considered to constitute an open class, meaning that postpositions can be (relatively) freely created on-line by the speaker. For example, an expression such as *jonkun avulla* 'with the help (of) someone' can well be considered a postposition although its relation to the noun *apu* 'help' is completely transparent.

5.2 English

For English, there is no need to go into detail about inflexional or derivational morphology. In this language, the morphology is very poor: adjectives are invariant, nouns have no marking for gender and their form only varies according to the number

(singular or plural), verbs usually have on form per tense, in other words their ending does not vary whatever the person and number of the subject, with the only exception of 3rd person singular in the present tense, which trigger an "s" suffix on the verb base. In addition, unlike Finnish, English only has one case which is visible on nouns, the genitive case expressing the possession, which is irrelevant in the framework of this project.

Therefore, we focus here only on two lexis-related issues: lexical unit formation (especially compounds) and collocations, very frequent in the environmental domain.

5.2.1 Compounds

In English there are two ways to connect two nouns together: with or without preposition. As it is the case in most Romance languages (French, Spanish, Portuguese, Italian, etc.), a noun can be connected to its complement with a preposition: *the quality **of** the air*, *the concentration **of** PM2.5*; *the density **of** particles*.
But like in Finnish or Germanic languages, it is also possible to create compounds – although in a different manner: English compounds are formed by a number of words (e.g., two nouns, as is often the case in the environmental discourse), with no syntactic connector, only by placing the head of the construction on the right: *air quality*, *PM2.5 concentration*, *particle density*, etc. Unlike Finnish or German, each noun keeps its lexical independence, i.e., both nouns are not "physically" merged into one noun, although the do behave like a single unit. In other words, the interpretation of what is a compound is semantic rather than morphological.
This option is very widely used and most of the time preferred to the use of a preposition in environmental information.

5.2.2 Collocations

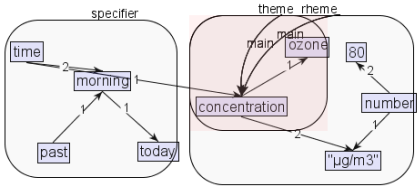

Figure 3: Semantic structure

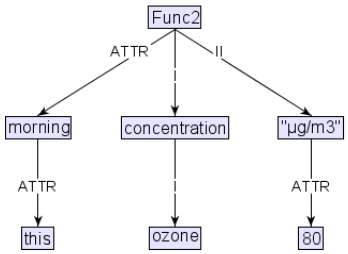

Figure 4: Deep-syntactic structure

Collocations, i.e. special cases of word cooccurrences, have to be handled by the system during the process of text generation. Let us take an example looking at a semantic structure –a predicate-argument structure containing all the meaning which has to be expressed in the final sentence- corresponding to the sentence *This morning, the concentration of ozone was 80 µg/m3*. In the semantic structure in Figure 3, one can notice that there is no node corresponding to the main verb *be*: actually, *be* in this case is a *support verb*, that is, a verb which does not have any meaning by itself but which is introduced only because a main verb is needed in order to built a correct sentence/clause.
In the Meaning-Text Theory, the linguistic framework used for the surface realization of sentences in this project, collocations have been formally described as *Lexical Functions* (LFs), which are functions in the same sense as mathematical functions: a particular function can be applied to a particular variable and return a particular value. For instance, in order to build the corresponding syntactic structure from the semantic

structure above, we need to introduce a main verb since none of the nodes of the semantic structure can be lexicalized as a verb (there is no verb corresponding to the meaning 'concentration', which is the main node of the structure). Thus, we are looking for a LF which will apply to the keyword *concentration* and return a verb that can take this keyword as a subject and its second argument (the concentration itself, *80 µg/m3*) as an object. Such an LF exists and is known as $Func_2$. This abstract lexical unit is introduced in the deep-syntactic structure as shown in Figure 4. Then, during the mapping between deep-syntax and surface-syntax, the values of the lexical functions are computed: the lexical function Func2 is computed and the value *be* is returned and introduced in the surface syntactic structure instead of the label of the function; cf. Figure 5.

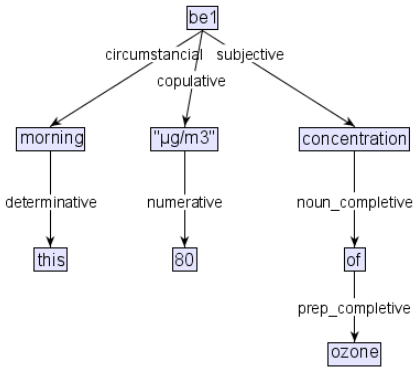

Figure 5: Surface-syntactic structure

Here are the main types of Lexical Functions which will be used in the domain of environmental information:
$Func_i(x)$: a LF which returns one or more verb(s) taking *x* as its subject and the *i*th argument of *x* as its object.
Ex: $Func_2$(concentration)=be: *the concentration **is** high*
$Oper_i(x)$: a LF which returns one or more verb(s) taking *x* as its object and the *i*th argument of *x* as its subject.
Ex: $Oper_1$(concentration)=have: *the ozone **has** a concentration of…*
Magn(x): a LF which returns one or more adverbial element(s) expressing the high degree of the keyword.
Ex: Magn(concentration)=high: ***high** concentrations of…*
AntiMagn(x): a LF which returns one or more adverbial element(s) expressing the low degree of the keyword.
Ex: AntiMagn(concentration)=high: ***low** concentrations of…*
Locin(x): a LF which returns one or more prepositional element(s) allowing the keyword to appear in adverbial paradigm.
Ex: Locin(area)=in: ***in** the metropolitan area, the traffic is dense…*

5.3 Swedish

5.3.1 Lexical issues: compounds and collocations

Compounding is a relatively frequent means for creating new lexical items. A typical Swedish compound consists of a noun or adjective as a base to which another noun or adjective, specifying the base, is added as a prefix, e.g. *samman-sättning* 'composition; lit. together-putting', *i huvud-sak* 'in general; lit. in head-thing', *gatu-damm* 'street dust', *stadsluften* 'city-air'. As the examples show, compounds can be either fully

lexicalized, i.e. fossilized combinations where the relation between the compound and its parts is not apparent, such as *sammansättning* and *huvudsak*. In both these cases the prefixed element appears in its nominative form. Other compounds reflect the historical evolution of the language with the prefixed element bearing case marking due to historical agreement relations between modifier and head. In *gatudamm*, for example, the ending in *-u* of the prefix, base form *gata*, reflects and old genitive ending whereas in *stadsluften* the *-s* is the genitive ending of today.
The same collocations which are presented for English (under 5.2.2) apply for Swedish as well:

- $Func_i(x)$: a LF which returns one or more verb(s) taking *x* as its subject and the *i*th argument of *x* as its object.
  Ex: $Func_2$(halt)=be: *halten* ***är*** *hög* 'concentration is high'
- $Oper_i(x)$: a LF which returns one or more verb(s) taking *x* as its object and the *i*th argument of *x* as its subject.
  Ex: $Oper_1$(concentration)=vara: *På Mannerheimvägen* ***är*** *halterna ungefär 6 gånger högre än…* 'On Mannerheimintie the concentrations are about six times higher…
- Magn(x): a LF which returns one or more adverbial element(s) expressing the high degree of the keyword.
  Ex: Magn(halt)=high: ***höga*** *halter…* 'High concentrations'
- AntiMagn(x): a LF which returns one or more adverbial element(s) expressing the low degree of the keyword.
  Ex: AntiMagn(halt)=high: ***låga*** *halter…* 'Low concentrations'
- Locin(x): a LF which returns one or more prepositional element(s) allowing the keyword to appear in adverbial paradigm.
  Ex: Locin(area)=in: ***i*** *huvudstadsområdet är trafiken livlig…* 'In the metropolitan area the traffic is dense'.

Negation
In general, negation is expressed only once in Swedish sentences. The negator *inte* usually follows the verb indicating a negative meaning, e.g. *Gatudamm är* ***inte*** *bara de stora städernas problem* 'Street dust is not only a problem of big cities'. Negative pronouns exist in Swedish (*ingen, inget, ingenting* 'nobody, noone') and when used impede the use of the negator, e.g. *Gatudamm är* ***inte*** *ett problem* vs. *Gatudamm är inget problem* 'Street dust is not a problem' vs. 'Streetdust is no problem'.

5.3.2 Word class characteristics
5.3.2.1 Nouns
Swedish nouns are divided into two classes, so called *common* and *neuter* **genders**. The gender determines the nouns definite form as well as that of any adjective modifier.
Plural is formed by adding a plural ending to the noun stem, but both stem and ending vary from noun to noun. Five classes or declensions are traditionally distinguished bearing endings in *-or, -er, -ar, -n* and no ending at all.
Definiteness is expressed in two ways. The definite article is a suffix while the indefinite article is a separate word derived from the numeral one (*en/ett*). The form of both articles depends on the gender, and is *en* form common nouns and *ett* for neuters, the definite suffix being *-(e)n* and *-(e)t* accordingly. E.g.

en *partikel*	***ett*** *utsläpp*
*Partikel****n***	*utsläpp****et***

The only remaining case ending is the genitive *-s*. It is currently regarded more as a clitic than as a pure case ending, since in complex nominal phrases it attaches not to the head noun but to the last element of the complex, e.g. *Mätstationen på Mannerheimvägen**s** index är högt* 'The station at Mannerheimvägen's index is high'.

5.3.2.2 Pronouns

Swedish has three series of personal pronouns as well as indefinite and demonstrative pronouns. The three series represent the nominative, accusative/dative and genitive case forms giving us different series of subject, oblique and possessive pronouns. No more details are given here since this information is not relevant to the air quality domain.

5.3.2.3 Verbs

Swedish verbs have lost all inflection for number and person, leaving the three tenses with one form each. Present tense always ends in *-r* and regular preterit forms in *-de*. Passive voice is expressed in two alternative ways, periphrastically by the use of auxiliary *bli* and synthetically by use of passive ending *-s* attached to the basic tense form. The synthetic passive is more frequently used than the English passive, while the periphrastic form is rare. Future tense does not exist in Swedish. For future reference either present tense or some periphrastic form can be used, such as *kommer att + INF.* 'come to + inf.', *ska* + INF 'shall + inf.'.

5.3.2.4 Adjectives

As stated above, modifying adjectives always follow their head in gender and number. For combination with neuter nouns the ending *-t* is added to the base form while adjectives modifying common nouns bare no ending. The plural ending is always *-a*.
Comparative and superlative forms of adjectives can also be formed synthetically or periphrastically. The comparative ending is *-(a)re* and the superlative ending *-st*, but some irregular forms exist (e.g. *stor – större – störst* 'big – bigger – biggest'). Periphrastic comparatives and superlatives are formed by using the comparative and superlative forms *mer* and *mest* ('more' and 'most') in front of the adjective.

5.3.2.5 Preps

Swedish prepositions are of two kinds: basic, monomorphemic prepositions which are semantically more vague and abstract and compound prepositions which generally exhibit a more specific, locative, temporal, causal or manner meaning. The most frequent locative prepositions are *i, på, vid, kring, över, under*, while *med, utan* and *för* are highly frequent notional prepositions. Common compounds include *på grund av* 'because of' and *i enlighet med* 'according to'.

5.3.2.6. Adverbs

Swedish adverbs consist of two types, invariant and derived adverbs. Adverbs of the first group generally express place and time, but also manner, while the derived adverbs are more varied. The formation of adverbs from adjectives is highly productive and implies the addition of adverbial suffix *-t* to the adjective. E.g. *snabb* 'fast' > *snabbt* 'rapidly'.

6 Conclusions

The analysis of the text and discourse structures on the one side and of the linguistic constructions in the environmental discourse in Finnish, English and Swedish on the other side has shown that the text generator in PESCaDO must be able to treat a rather wide range of linguistic phenomena – which implies that a template-based generator as used for linguistically only little varying domains cannot be the solution in PESCaDO: the effort to implement all possible templates and to ensure a sufficient flexibility for further extensions to other related domains would be too high. A genuine multilingual text generator that starts from abstract input structures (to be provided by the content selection module) and that produces a high quality textual output by successively mapping intermediate structures (from more abstract to more concrete) is needed. For this purpose, UPF will extend its rule-based generator. In parallel, Finnish, English and Swedish corpora will be annotated with all necessary types of linguistic structures in order to facilitate machine learning-based text generation. This is in order to ensure the highest scalability and best performance possible.

References

Mann, W.C. and S. Thompson. 1988. Rhetorical Structure Theory: A theory of text organization. In L. Polanyi (ed.) *The Structure of Discourse*. Norwood: Ablex.

Mel'čuk, I. 1988. *Dependency Syntax: Theory and Practice*. Albany: SUNY Press

Mel'čuk, I. 1996. Lexical Functions: A tool for the description of lexical relations in the lexicon. In L. Wanner (ed.) *Lexical Functions in Lexicography and Natural Language Processing*. Amsterdam: Benjamins Academic Publishers, pp.37-102.