Looking back to move forward.
Challenges related to deceitful parallel texts and slippery terms

Kai Innselset, Marita Kristiansen & Kari Øvsthus

Department of Professional and Intercultural Communication

Norwegian School of Economics and Business Administration (NHH)

Summary
In this article we will address some pitfalls and shortcomings related to semi-automatic term extraction and an uncritical reliance on translational equivalents in parallel texts. After a short description of the KB-N project and its corpus material and semi-automatic term extraction tool, we shall go on to give examples of faulty equivalents in translated texts resulting from adaptations to an external context. Next, succeeding a comparison of different views of "termhood", we shall give examples of how even highly domain-focal terms can go unnoticed by an extraction tool. Finally, we shall present two case studies: verbalisation and clipping. The former study shows how semi-automatic extraction may fail completely and manual extraction be impeded owing to linguistic choices made by a translator in representing certain key concepts. The latter study shows how the absence of full-term realisations in a text can cause confusion as to which concepts are involved if the text is handled exclusively by an extraction tool. The overall conclusion is that there are characteristics of texts which are beyond the control of regular semi-automatic extraction tools, and that human intervention is indispensable.

1.Introduction

In the not too distant past, the only means a terminologist had when extracting terms was his marker pen which easily highlighted the linguistic units perceived as denoting special content relevant to a specific domain. With the development of language technology tools, the day-to-day activities of the terminologist may seem much easier to handle.

The development of useful language technology tools was the main focus of the Norwegian Research Council's research programme KUNSTI, from which the 3-year project KB-N (Knowledgebank Norway) at the Department of Professional and Intercultural Communication, Norwegian School of Economics and Business Administration (NHH) has received funding.

The aim of the KB-N project, headed by Professor Magnar Brekke, has been to establish a concept-oriented text and term-based knowledge management system which includes language technology applications for use primarily within translation, documentation and publishing. In particular, the system includes 15 economic-administrative domains, such as *financial accounting*, *marketing* and *economics*. These domains are represented in a termbase consisting of approximately 8,500 term records and a textbase, consisting of some 1.8 million running words.

Although the overall goal of the KB-N project has been to develop language technology tools, anyone who has been involved in such research knows that it is necessary to collect

data from which to extract information to test the tools that are to be developed. In this article we will address one of our major sources of data, namely our solid collection of parallel texts. Our parallel corpus, consisting of texts in both Norwegian and English, constitutes almost half of the texts included in the textbase.

Furthermore, we will discuss some methodological challenges we have come across when extracting terminological information from the parallel texts to be registered in our termbase. Thus we will take a closer look at how far the modernisation of the terminologist's methods for extracting terms has taken us in our quest for easily-recordable terminological units.

In the article we will first outline the main framework for the extraction of equivalents, including a brief discussion of our corpus material and the semi-automatic term extraction tool that has been developed. Next, we will present the theoretical basis that has been guiding for some of the choices we have made when building our knowledge base from our parallel corpus. Finally, we will provide some examples that illustrate the shortcomings of relying too heavily on parallel texts for the extraction and compilation of terminology.

2. A framework for extracting equivalents

2.1 Domain-specific terminology

An underlying methodological principle which may, as we will illustrate later, cause errors in connection with semi-automatic term extraction is our demand for a domain-specific presentation of data in our knowledge base. It is a typical characteristic of the economic-administrative domains that the terminology resembles the words of everyday general language. Therefore the number of polysemes may appear to be high if the terms are not assigned to their special domains to which they belong. The project has therefore aimed at assigning domain markers to the term records. This entails that a lexical unit may appear in several term records, however, then with a unique domain marker (see Kristiansen 2005 for a more detailed discussion).

2.2 Corpus material

One part of our corpus consists of texts that are single, but domain-specific units that exist in both English and Norwegian. Such corpora are frequently described as *comparable corpora*, i.e. similar texts in more than one language or variety. Such texts allow us to analyse the terminology used in similar circumstances of communication, without the inevitable distortion introduced by the translations of a parallel corpus. These texts belong to the same domains and their content and level of technicality correspond.

One example of such comparable corpus texts may be speeches given by the Norwegian Central Bank Governor and the US Federal Reserve Chairman, respectively. Other such corresponding texts are introductory textbooks aimed at the same group readers, i.e. students at university level who study the domain in question. One example here may be Norwegian and English textbooks in economics, respectively.

Although the comparable corpus makes up a substantial part of our textbase, a very important feature of our project has been our strong reliance on parallel texts. These have been used to establish a solid framework for extracting equivalent terms in English and Norwegian in the various economic-administrative domains included in our study. Such texts have also been needed as input to develop our automatic term extraction tool which we will return to below.

A *parallel corpus* is a collection of texts, each of which is translated into one or more other languages than the original. In our study, only two languages have been involved, namely English and Norwegian. The direction of the translations has not been constant, although the majority of texts in the parallel corpus have been translated from English to Norwegian.

There seems to be an unwritten rule among scholars that a corpus should comprise a certain amount of text and word samples to qualify as a scientifically sound material. The London-Lund corpus, for example comprises 435,000 words, and consists of 87 texts of 5,000-word samples (Stubbs 1996: xvii). For use in a terminological analysis this criterion is somewhat naive. Selecting a certain number of texts containing a specified number of words may not offer any useful material at all (Bergenholtz 1996, Picht 1996, Kristiansen 2004). In order to disclose an already existing terminology, i.e. the terms of the field, completeness must be an overall goal. Since a domain will be developing dynamically, this will of course have to be a goal that can only be partly met.

2.3 Semi-automatic term extraction in KB-N

To ease the term extraction process, we wanted to develop a semi-automatic tool for extracting Norwegian terms since no such tool was available in our native language.

Manual term extraction is said to be reliable and result in "the correct term list", but at the cost of being labour intensive, time consuming and dependent on individual choices. Semi-automatic term extraction, i.e. automatic generation of lists of candidate terms for subsequent evaluation, is imbued with errors such as overgeneration (noise) and undergeneration (silence).

For semi-automatic extraction, there are two main strategies: a linguistic approach and a statistical approach. A statistical approach takes the occurrence ratio of an item into account. Hybrid extraction tools, which are very common, combine the two approaches. Such tools take the occurrence ratio of an item into account in addition to its linguistic form.

Extraction relying on linguistic criteria attempts to identify terms by matching words and word combinations in the text against a predefined list of specific linguistic patterns for term formation. A list of matches is then automatically generated, a number of which will be automatically discarded for statistical reasons if a hybrid tool is used. The words or word combinations in this list are known as candidate terms.

As a result of their limitations, term-extraction systems add word combinations to their list that a terminologist would not consider to be terms. These "non-terms" are referred to as "noise". The amount of noise generated by a term-extraction system is in direct relation with the number of terminological patterns it attempts to extract. The system must try to strike a delicate balance between including the highest number of relevant term formation patterns possible and keeping noise to a minimum (Love 2000).

The KB-N system uses a hybrid extraction method incorporating both a linguistic and a statistical approach. The primary module of the linguistic component is an algorithm consisting of a list of major noun phrase patterns relevant for term extraction. It is nothing more than a structural description of a small subset of typical noun phrases. A description of this subset is found in Øvsthus et al., 2005.

This algorithm is rather minimalistic and consequently not very noise-producing. For further noise reduction we have included a stop list for adjectives as an extension of the linguistic component. It contains approx. 300 adjectives which we assume will seldom or never occur as modifiers in multiword terms.

The statistical component of the KB-N extraction module ensures even further noise reduction. It makes use of a "weirdness filter" which suppresses strings which occur more frequently in an LGP corpus than in the LSP text under consideration, but which otherwise satisfy the structural criteria of the linguistic filter. As reference corpus we use the Norwegian newspaper corpus of approximately 350 million words (http://avis.uib.no/english.page).

When a term-extraction system fails to identify terminological units in the text, the error is classified as "silence". Important causes of silence are overly-stringent term candidate identification rules in the linguistic filter and weaknesses relating to the noise reduction mechanisms mentioned above. In general, automated term-extraction programs are designed in such a way as to reduce silence to a minimum (Love 2000). We cannot claim this for the KB-N extraction tool. Our focus has been very much on noise reduction. Thus our linguistic algorithm does not support the extraction of postmodified noun phrases, notably NP+PP, a highly relevant term pattern. Noise production associated with this pattern relates among other things to problems with syntactical disambiguation. A PP can appear as a phrase-internal modifier, but also as an adverbial sentence element in its own right immediately following an NP. To compensate for the silence resulting from the non-implementation of the NP+PP pattern, we rely on picking up such terms indirectly from the context. This of course presupposes that at least one of the NPs in the pattern has made it to the candidate list.

3. Exemplifications and case studies

In the following we will discuss some of the shortcomings we have experienced when developing our KB-N knowledge base. Firstly, we will discuss the obvious source of incorrect terminology due to context-adapted terminology. Secondly, we will address the use of verbalisations as textual realisations of concepts. We will present a special case study to illustrate our points. Finally, we will discuss the problems caused by consistent clipping, i.e. the fact that some concepts are never realised by their full terms, by means of a second case study. In our discussion we will use single quotation marks to express terms and italics to express concepts.

3.1 The illusion of equivalents in texts with context-adapted terminology

One possible source of error when basing a pair of terms on parallel texts is context-adapted choices of terms in the target language, made by the translator. The following examples will illustrate this point.

In 2004, *Enterprise Risk Management – Integrated Framework* by the Committee of Sponsoring Organizations of the Treadway Commission was published. This was translated into Norwegian the year after. The Norwegian translation includes a term list, where the translators account for some of the choices made in understanding with their employers, reflecting both conceptual and terminological adaptations to fit the translations to a Norwegian corporate context.

A term frequently used in the English source text is the term ‘company’, which implies a type of business firm with limited liability. There is an equivalent term in Norwegian, ‘selskap’, but this has not been used as the prevalent term in the Norwegian target text. In stead, the Norwegian term ‘foretak’ is used, which is equivalent to English ‘enterprise’. The reason for this choice was that using the term ‘foretak’, would make participants in a wider range of organisation types feel that the principles described in the publication could be relevant to their organisations.

In the same publication the term ‘entity’ is widely used to refer to the object of enterprise risk management. In the Norwegian target text, this has been translated by the Norwegian term ‘virksomhet’, which means ‘activity’, ‘business’, or even ‘enterprise’, and is a narrower concept than ‘entity’. There are Norwegian equivalents to ‘entity’, such as ‘entitet’ or ‘objekt’, but these would seem somewhat grand and even unidiomatic if used in the Norwegian text.

If pairs of terms, such as ‘company’ – ‘foretak’ and ‘entity’ – ‘virksomhet’ are perceived as general equivalents because they are found as equivalents in a set of parallel texts, this could have unfortunate consequences. This shows that the knowledge and attention of a terminologist is certainly needed to avoid errors easily made by uncritically accepting translational equivalents. The examples also show that our domain focus would not eliminate such possible errors, since the texts are true parallel texts.

3.2 Untypical terms

Whether excerpting candidate terms manually, or programming an automatic module to find them, it is necessary to have a clear view of what a term is. Experts are not in complete agreement with regard to this question.

Table 1 (adapted from Øvsthus, doctoral dissertation, forthcoming) provides an overview based on discussions of the concept *term*, where different views are taken into account. The arrow to the left represents the degree of “termhood” or “technicality”, and the other columns of the table represent different views on the concept ‘term’, organised so that the different categorisation systems or continua correspond. We would like to point out that it is not our intention to suggest that “termhood” can be measured in any exact manner, but merely to indicate direction as we move between categories or along continua.

Degree of termhood	Sager et al. (1980)	Hoffmann (1985)	Brekke (2006)	Role of context	Proto-typicality	Domain-focality
↑	Terms	Subject-specific vocabulary	Terms unique to a specific subdomain	↓	↑	↑
			Homographs covering terms denoting different concepts in different domains			
	Words	Non subject-specific specialised vocabulary	General academic and scientific terms and research-oriented terms			
		General vocabulary				

Table 1

Sager et al. state that only items with special reference within a domain are terms of that domain, and that items which function in general reference in more domains are simply "words" (1980:75). Sager et al. do not state specifically how items which can function in more than one domain, but with special reference within each one of them, should be viewed. The view of Sager et al. is represented by a division into two categories as shown in the second column of table 1.

Hoffmann (1985:126f.) is willing to admit part of what Sager et al. dismiss as 'words' into the ranks of terms, namely what he refers to as "allgemeinwissenschaftlicher Wortschatz", meaning vocabulary which is frequently found in special language texts, but does not belong to one specific domain alone. He makes a distinction between this category being the "non subject-specific specialised vocabulary" and "general vocabulary", which is reflected in a division into three categories in the third column.

A similar category to Hoffmann's "allgemeinwissenschaftlicher Wortschatz" is shown under Brekke (2006), where his "general academic and scientific terms" and "research-oriented terms" have been gathered into one category. In Brekke "homographs covering terms denoting different concepts in different domains" have been specifically mentioned, and a separate category for these has therefore been shown. This is to separate these from the strictly subject-specific terms.

In the fifth column the role of context for each category is shown. The downwards-pointing arrow indicates that the higher up in the "hierarchy" of terms a given item belongs, the less need for context to establish its meaning. For a term which denotes different concepts in

different domains, context is needed to make the right decision about which one of the possible concepts the term denotes in a given case. In the case of terms denoting seemingly identical concepts within different domains, context may be needed to establish the precise meaning of the term.

As for prototypicality, the arrow in the sixth column reflects the assumption that terms from the topmost category will have the greatest probability of being perceived as a prototypical term, i.e. a good example of a term. This must be so since all authors and researchers agree that items which belong in the topmost categories are terms. When moving further down, there is considerable disagreement, and the further down, the less inclination to grant term status to items belonging in the relevant category.

Finally, Brekke (2006) initially assumes that only terms unique to a specific domain can be considered domain-focal terms. Subsequent discussions, reveal, however, that it cannot be excluded that items from other categories could be perceived as domain-focal. The upwards-pointing arrow in the seventh column nevertheless indicates that there is a greater likelihood that a term from the topmost category will be domain-focal than a term from the bottom category. This may not be true for all domains, indeed we believe that many economic-administrative domains have focal terminology from other categories than the topmost, this being related to the assumption mentioned before, that many terms from this domain resemble words from general language.

This may give cause to challenges when the task is to identify terminology from a domain. If the task is to be performed by a semi-automatic tool, such as the one described in section 2.3 above, candidate terms which resemble general language words may be "caught" in the "weirdness filter" because they are not "weird" enough. An example of a term which would run such a risk is the Norwegian term 'balanse' (in English: 'balance sheet'). This is an extremely focal term within the domain of accounting, which has a general language polyseme. This particular term would probably not cause any difficulties to a manually excerpting terminologist (with sufficient domain expertise), if the item could be disambiguated by context.

Slightly more devious are items belonging to Hoffmann's "non subject-specific specialised vocabulary" or Brekke's "general academic and scientific terms and research-oriented terms". Examples from this category are 'analysis' and 'report'. These items may also run a risk of not making it through the "weirdness filter" of an automatic tool, since they may occur with a certain frequency in an LGP corpus. To the terminologist, ambiguity is not the main problem with this category. Here the terminologist will have to ask herself: Are these strings part of the terminology of the domain in question, or are they general language words, but belonging to an academic style of writing? Where do I draw the line? It could be said that this is a problem related to collecting a terminology for a domain by term extraction, as opposed to building a terminology on the basis of a concept system. It is a problem which arises when we are trying to judge the terms by "face value", rather than looking for them on the basis of concepts.

3.3 Textual realisation of concepts

Linguistic representations of concepts are not restricted to terms. Other possible realisations include paraphrases, LSP phrases, definitions and explanations (Picht 2007). When textual realisations of concepts rely heavily on these alternative representations, the term extraction

module will often draw a blank. This of course also applies whenever a text tends to realise concepts systematically by means of linguistic term patterns which are not implemented in the linguistic filter, notably verbs and verb phrases, and whenever concepts are systematically realised by clipped terms which otherwise conform to an implemented pattern, but never appear in their full form. In the following we shall study these two phenomena under the headings "**verbalisation**" and "**clipping".**

The two phenomena will be studied with reference to the texts "Positioning and differentiating the market offering through the product life cycle" (chapter 11 in the introductory text book *Marketing Management* by Kotler (2003), and its Norwegian translation entitled "Å posisjonere og differensiere markedstilbudet gjennom produktets levetid" (Kotler 2004). Each text amounts to 36 pages. Verbalisation is a prominent feature of the Norwegian text, whereas clipping occurs equally frequently in both texts.

3.3.1 Verbalisation

A concept system covering the content of the entire chapter was set up based on the English original. This was a fairly simple task. The text being from an introductory text book covering a relatively non-technical subject field, the content was readily understood. The terminology was presented and discussed in an orderly fashion, often with sets of subordinate terms presented as list headed by their respective superordinate terms. Much of the concept system could therefore be set up sequentially as we proceeded page by page through the text.

In order to illustrate the consequences of verbalisation in the Norwegian version for term extraction (both semi-automatic and manual) and subsequent term registration in the terminological database, we shall focus on some major concepts relating to *positioning*. The concept of *positioning* is defined as follows: "The act of designing the company's offering and image to occupy a distinctive place in the mind of the target market" (Kotler 2003:308). The relevant concept system fragment based on the English version can be presented as in figure 1 below:

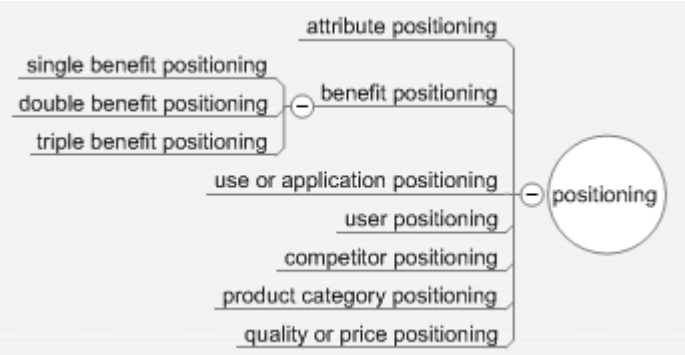

Figure 1

We shall anticipate the course of events by the following statement: Semi-automatic term extraction from the Norwegian version would draw a complete blank for the entire concept system above. Moreover, even manual extraction would possibly have missed out on some concepts, viz. the Norwegian "realisations" of 'single', 'double' and 'triple benefit positioning'. These concepts will be discussed below.

All the English terms in this system are premodified noun phrases. They are all matched in the Norwegian translation by postmodified verb phrases, which are naturally not picked up by the extraction module (ref. table 2 below). As for determinacy, the solution selected by the translators is superior. They can be considered unpacked translational equivalents of the English terms, showing unambiguously the adverbial nature of the postmodification.

The English terms, all on a "classical" term form, are indeterminate, some more than others. Andersen (2000:21) remarks that a typical feature of packing of arguments in nominalisations is that distinctions which are relatively clearly coded by verbal arguments (e.g. the distinction between direct objects and adverbials) tend to get lost in packed constructions. It is a general fact that the more you pack expressions, the less determinacy you get (Andersen 2007). Thus the packed premodified English terms with the deverbal "positioning" as head are all ambiguous as to whether the premodification should be interpreted as a derived object or a derived adverbial. Andersen (ibid.), describing Norwegian compound deverbal nouns, states that when the first element is animate (especially when it refers to a human being) the interpretation as derived subject or derived direct object is genuinely underdetermined syntactically. This observation applies to two of our terms: 'competitor positioning' and 'user positioning' They are both ambiguous not only as to the interpretation of the premodification as direct object or adverbial, but also as to an interpretation as derived subject.

Indeterminate as the English terms may be when considered in isolation stripped of any disambiguating clues, they are certainly not ambiguous in the text. The underlying concepts are fully explained, and sometimes even nearly defined, as the corresponding terms are mentioned, i.e. they are disambiguated by the co-text.

As for the Norwegian translational equivalents, these are fully determined. Being verb phrases, however, they are not suited as equivalents to the English terms in a terminological database. It so happens that our text considers each of the underlying concepts only once, thus depriving us of any potential anaphoric realisations in the form of nominalisations with the verb phrases as antecedents. Norman (2003:114) defines an anaphor as "a word or phrase that refers back to one or more words or phrases occurring earlier in the same text". To establish proper equivalents, we shall therefore have to nominalise these verb phrases ourselves. The alternative would be to re-translate the English terms following the same pattern of premodification. The latter alternative is obviously not a good idea as this would result in the same indeterminacy as discussed above concerning the English terms.

It seems imperative then to resort to the less packed and more determinate construction of deverbal noun with postmodifying prepositional phrase, which, thanks to the preposition "etter" ("by/according to/with respect to"), unambiguously signals a derived adverbial, though indeterminate as to time and manner. In fact, representing these concepts by means of terms with a premodifying element would not, neither for English nor Norwegian, be neutral as to the interpretation of the inherent ambiguity. In isolation the premodification of such terms would probably (for a non-specialist) be interpreted as a derived direct object when the premodifying element is inanimate, and either derived subject or object when the premodifying element is animate (especially when denoting human beings). The terms from the English original, their translation into Norwegian, and the derived Norwegian terms are presented in table 2 (next page).

English original	Norwegian translation	Derived Norwegian term
attribute positioning	å posisjonere etter egenskap	posisjonering etter egenskap
benefit positioning	å posisjonere etter fordel	posisjonering etter fordel
use or application positioning	å posisjonere etter bruk og anvendelse	posisjonering etter bruk og anvendelse
user positioning	å posisjonere etter bruker	posisjonering etter bruker
competitor positioning	å posisjonere etter konkurrent	posisjonering etter konkurrent
product category positioning	å posisjonere etter produktkategorier	posisjonering etter produktkategorier
quality or price positioning	å posisjonere etter kvalitet og pris	posisjonering etter kvalitet og pris

Table 2

Summing up, manual term extraction from the English original and a comparison with the Norwegian version have resulted in a complete subset of English terms relating to *positioning* with their respective derived Norwegian equivalents, which are all nominalised versions of the verb phrases in the Norwegian text. All the derived terms have kept the high level of determinacy of the corresponding verb phrases, and are therefore superior to the English terms, which are on a lower level of determinacy.

Any attempt to uncover these concepts by applying our automatic extraction module to the Norwegian text would have been futile. Like many other existing extraction modules, the KB-N extraction module does not support extraction of verbs. Implementing verbs and verb phrases as a valid term pattern in the module could only be done at the expense of creating a lot of noise in the candidate term list.

A relevant question is who is cheating us for capturing the textual realisations of these concepts. Is it we ourselves, for failing to implement verbs as valid term structures and thus creating what is known as “silence”, or the translator, for having a predilection for verbal expressions?

Maybe we can throw a little light on this question by considering what would have happened had we tried to identify the concepts by manually going through the Norwegian text. Let us concentrate on the Norwegian textual realisations corresponding to the concepts *single-*, *double-*, and *triple-benefit positioning*. To get our point over more clearly, we have also provided the entire contexts for both languages in the table below. As opposed to the English contexts, where the concepts leap to the eye realised by a familiar term structure, the Norwegian contexts provide no such clue. We should be hard put to capture the three concepts; there is very little to indicate that a classification is intended. This impression is strengthened by the vacillation between the verbs “kjøre” and “posisjonere”.

English original	**English context**	**Norwegian translation**	**Norwegian context**	**Derived Norwegian term**
single-benefit positioning	Not everyone agrees that single-benefit positioning is always best.	kjøre bare én fordel	Men ikke alle er enige i at det alltid er best å kjøre bare én fordel.	posisjonering etter én fordel
double-benefit positioning	Double-benefit positioning may be more distinctive.	posisjonere seg med to fordeler	Man kan skille seg ut ved å posisjonere seg med to fordeler.	posisjonering etter to fordeler
triple-benefit positioning	There are even cases of triple-benefit positioning.	posisjonere seg med tre fordeler	Det finnes til og med eksempler på at man med hell har posisjonert seg med tre fordeler.	posisjonering etter tre fordeler

Table 3

Why the translator has landed on a verbal style is hard to say. A possible explanation may be that the text is relatively non-technical, a fact which might have triggered the desire to adopt the verbal style which we are encouraged to use in LGP texts. An exaggerated and unnecessary use of deverbal nouns instead of the underlying verb is not considered good language. There is even a term for this violation of received LGP style in Norwegian, namely "substantivsyke" ("noun disease"). A felt necessity to use verbal expressions in order to achieve designations superior to the English terms with respect to determinacy can hardly explain the verbal style, as this could equally well have been achieved by using a deverbal noun with prepositional postmodification.

3.3.2 Clipping

We shall use the term clipping in a loose sense to denote any reduction of lexical elements in multiword/compound terms denoting a particular concept. Clipping is relevant for anaphoric reference. Rogers (2007) claims that the repeated use of the full term could, for textual reasons, be disorienting for the reader who assumes that he/she is being given new information. Repeated use of the full term would be a case of overspecification. She continues:

> Another way of viewing this is that language use tends to a certain economy of expression, although this is mitigated by the need for precision in the use of terms. Appropriateness acts as the arbiter between economy and precision [...] It is nevertheless not possible to predict exactly where clipping will take place and how (ibid.:30).

We shall also use the term clipping indiscriminately to denote terms which are reduced versions of longer (hypothetical) terms which are not realised in the text, but will have to be derived/constructed before they can profitably be included in their respective term records and placed in a concept system. Thus we shall use the term clipping in a wide sense covering more than pure anaphoric reference.

From the point of view of a terminologist who aims at extracting terminology from a text, semi-automatically or manually, clipping, though in many cases an inevitable phenomenon, can present an obstacle. Consistent use of full terms throughout the entire text would be preferable. This would imply the use of the full term at first mention and subsequent straight repetition of the full term as anaphoric reference.

In order to illustrate the phenomenon clipping, we shall concentrate on the concept system fragment shown below taken from the chapter fragment "market evolution" (Kotler 2003:399ff.), which describes the various stages of market evolution. As opposed to the concept system used for illustrating verbalisation, where the problems relating to term extraction and equivalence were caused by the Norwegian translation alone, clipping necessitated full-term derivation for both language versions. The concept system in figure 2 below is therefore the end product following the derivation procedures.

Figure 2

The chapter fragment macro structure for the two language versions are:

Market evolution	Markedsutvikling
stages in market evolution	markedsutviklingens faser
emergence	tilblivelsesfasen
growth	vekstfasen
maturity	modningsfasen
decline	tilbakegangsfasen

This macro structure unequivocally points to a hierarchy of concepts. The textual realisations of the four subordinate concepts in the English version are clipped at both ends. Rogers (2007:23) uses the term "double-ended clipping" to denote "terminological toggling" involving the alternate clipping of the left and the right element of a complex term. We shall therefore label the English terms above "simultaneously double-ended-clipped", this seen in relation to the full terms with a grammatical head to the right and an extra modifier to the left. The Norwegian realisations in the macro-structure are only left-clipped and thus less underspecified. In both language versions simultaneous double-ended clipping is used for the first realisations in the running text: "Like products, markets evolve through four stages: emergence, growth, maturity, and decline" (Kotler 2003:340) and "I likhet med produkter gjennomløper markedene fire faser: frembrudd, vekst, modning og tilbakegang" (Kotler 2004:279).

Table 4 (next page) gives a summary of the realisations in order of mention in the texts:

English	**Norwegian**
emergence	frembrudd
emergence	tilblivelsesfase
Emergence stage	tilblivelsesfase
growth	vekst
growth	vekstfase
Market growth stage	vekstfase
maturity	modning
maturity	modningsfase
Maturity stage	modningsfase
decline	tilbakegang
decline	tilbakegangsfase
Decline stage	tilbakegangsfase

Table 4

The only full-term realisation, both languages considered is 'market growth stage', which appears in the context "If the product sells well, new firms will enter the market, ushering in a market-growth stage" (Kotler 2003:340). The corresponding Norwegian realisation is left-clipped: "Går salget av det nye produktet godt, vil nye bedrifter strømme til, og markedet er snart inne i sin vekstfase" (Kotler 2004:280). The choice of the clipped term here is in fact the only sensible alternative considering the juxtaposition of words in the clause. Using the "full" Norwegian term leaving the clause structure otherwise unchanged would be misleading beyond pure overspecification. As the clause explicitly signals the genitival relationship between 'marked' and 'vekstfase', the use of the full term would open for the interpretation that there were more than one type of 'vekstfase' pertaining to the market: '*.....markedet er snart inne i sin markedsvekstfase'. This is seen perhaps even more clearly if we attempt a genitival phrase structure: '*markedets markedsvekstfase'.

In terms of anaphoric reference, the various realisations of both the English and the Norwegian version appear in the wrong order, so to speak, i.e. from clipped to expanded realisations. Only one realisation reaches the level of full term and then in its final mention, namely the 'market growth stage'. This is contrary to the more usual techniques for anaphoric reference, one of which is what Norman has termed "reductive head-repetition", defined as "repetition of the head of a nominal group, but with elimination of some or all of the modifiers present in the antecedent" (Norman 2003:119).

The maturity stage turns out to be divided into two consecutive substages realised by the unclipped terms 'market-fragmentation stage' and 'market-consolidation stage' in a figure in the margin, which would not have been included in any electronic text. The same applies to the Norwegian version, which has 'markedsfragmenteringsstadiet' and 'markedskonsoli-deringsstadiet', respectively.

Table 5 (next page) shows the fullest concept realisations observed in the texts together with derived terms.

English original	Derived English term	Norwegian translation	Derived Norwegian term
Emergence stage	Market emergence stage	tilblivelsesfase	markedstilblivelsesfase
Market-growth stage		vekstfase	markedsvekstfase
Maturity stage	Market maturity stage	modningsfase	markedsmodningsfase
(in figure only) Market-fragmentation stage		(in figure only) markedsfragmenterings-stadiet	
(in figure only) Market-consolidation stage		(in figure only) markedskonsoliderings-stadiet	
Decline stage	Market decline stage	tilbakegangsfase	markedstilbakegangsfase

Table 5

Matters are further complicated by the fact that markets are not alone in passing through a number of distinct stages. Our text also describes the various stages for *products* and *fashions*. These are, like the majority of textual realisations for the market evolution stages, underspecified in both language versions in the sense that no full term is ever realised. The fullest terms are left-clipped realisations. This again makes it necessary to derive sets of full terms before inclusion in the terminological database. Like the treatment of the concepts relating to the market evolution stages, the order of mention in the texts starts with the most clipped realisations.

Like markets, a product passes through four stages, three of which are identical to the market evolution stages. One stage, *maturity*, is subdivided into three phases. The translation of 'stage' is now 'stadium' as opposed to 'fase' for the various market evolution stages, motivated perhaps by the use of 'phases' for the subdivisions of *maturity*. The compounds involving 'fase' and 'stadium' for markets and products, respectively, will be construed as pair-wise synonyms, the suspicion being confirmed by the consistent use of 'stage' in the English original. Likewise, the concept of *fashion* passes through four stages, one of which is *decline*. Here 'fase' is used once again as the equivalent Norwegian term.

To sum up, consistent clipping, i.e. the total lack of full-term realisations, present few problems for manual term extraction based on parallel reading of the texts. The challenge is that non-realised full terms will have to be derived from the occurring clipped terms. The fact that say e.g. the double-ended clipped term 'decline', with its Norwegian counterpart 'tilbakegang', represents three different concepts whose non-realised full terms are/will be, respectively 'market decline stage', 'product life-cycle decline stage' and 'fashion decline stage', and their Norwegian counterparts, is really not much of a problem. The context, the immediate as well as the wider, offers all necessary clues for "disambiguating" any occurrence of the triple-ambiguous "decline". Should we ever need an excuse for expanding such clipped terms, and a pattern for expansion, we need only refer to an attested full-term occurrence: e.g. 'market growth stage'.

For semi-automatic extraction consistent clipping presents serious problems. First of all, as opposed to manual extraction where the terminologist can use physical paper copies of the documents, semi-automatic extraction relies solely on the use of electronic texts. Any graphics in the form of figures and tables are removed, i.e. only running text is considered. Thus full-term realisations in figures and tables will be physically lost, such as 'markedsfragmenteringsstadiet' and 'markedskonsolideringsstadiet', together with the corresponding full terms in the English version.

For clipped concept realisations actually occurring in an electronic text, the first challenge is of course to be accepted as a candidate term. Like any other concept realisations they will not appear in the candidate list unless they satisfy the term candidate criteria of the linguistic filter and the weirdness filter. As for the realisations we consider here they all qualify as far as the linguistic filter is concerned, but the weirdness filter poses a serious threat. Double-ended clipped terms are likely to be suppressed. Terms like 'vekst', 'modning' and 'tilbakegang' are after all very general words with a high frequency in the LGP reference corpus. The left-clipped terms will have a greater chance of being picked up. The most interesting information in terms of definitions and even statements concerning the subdivisions of market evolution stages and product life-cycle stages is, however, linked to the double-ended clipped terms. The terminologist will therefore have to take pot luck as to whether he or she will be able to spot that the different compounds involving 'fase' and 'stadium' that may turn up in the candidate list pertain to different concepts such as markets, products and fashions.

4. Concluding remarks

As we have demonstrated in this article, concepts in texts lead dangerous lives. They are wholly dependent for their recognition and registration on whether their textual realisations are perceived and captured by manual scrutiny or man-made semi-automatic extraction procedures.

An author of an LSP text has the reader in mind. He may choose to convey his message including the domain's concepts in various ways. Thus concepts may appear in many guises, not only as fully motivated terms.

A terminologist has a different focus from that of the general reader of a text. His aim is to look for concepts in all their various linguistic guises, fit them into a concept system showing the various relations between them, and represent them with their names, i.e. their terms, in the concept system and in a terminological database.

In this article we have discussed how our reliance on corpora, especially used in combination with semi-automatic term extraction from Norwegian texts with accompanying equivalent-detection from parallel English versions, has led to a higher number of undetected concepts and errors than what would have been the case if the more traditional terminological approach of manual term extraction had been used. This should come as no surprise. The role of semi-automatic extraction is limited to presenting the terminologist with a list of candidate terms for further scrutiny, each candidate owing its appearance in the list to its conformance to a predefined linguistic pattern and to its required frequency of occurrence in the text as compared to its occurrences in a large reference corpus. That is a long way to go for a term in the KB-N extraction module, which we have deliberately programmed to produce tidy and

relatively noise-free candidate lists. This is achieved mainly by setting a high threshold level for the occurrence ratio.

Our case studies show that we are clearly cheated for capturing the textual realisations of a number of the concepts we have analysed. This is especially true for semi-automatic extraction, but it may also be true, though to a lesser extent, for manual extraction.

In some instances it may be the translator who is cheating us. If that is the case, a timely question is why this is done. One obvious reason may be that at least in Norwegian, a comprehensive terminology may be lacking. Other possible explanations include cultural adaptations, such as discussed in section 3.1, and verbalisations used for pedagogic reasons or to increase determinacy.

Such moves made by the translator may increase the pedagogic value of a given text, in that it may appear to be easier to understand. On the other hand, the lack of proper terms in textbooks may represent a challenge for students of a given domain, since nominalisations would give the students a means to relate the new knowledge to already existing one. By hiding the knowledge in running text without such easily recognisable expressions, it is possible that the learning outcome will be lower than perhaps what could be expected given the effort put into the reading.

The lack of easily recognisable expressions has also been a challenge to us when building our knowledge base. Developing our semi-automatic term extraction tool has given us valuable insight and a good basis for addressing the challenges still ahead. Having a language technology tool has proved to be a useful means in the process, but there seems to be no way to disregard the human mind if we are to develop a high-quality product – and maybe that is just as well.

5. References

Andersen, Øivin (2000) Analyse av komplekse nominaliserte termer. In: Nuopponen, Anita /Toft, Bertha / Myking, Johan (eds.) *I terminologins tjänst. Festskrift för Heribert Picht på 60-årsdagen*. Vasa: University of Vaasa. 2-24.

Andersen, Øivin (2007) Indeterminacy, context, economy and well-formedness in specialist communication. In: Bassey, Edem Antia (ed.) *Indeterminacy in terminology and LSP*. Amsterdam/Philadelphia: John Benjamins. 3-14.

Bergenholtz, Henning (1996) Korpusbaseret leksikografi. *LexicoNordica,* 3. 5-17.

Brekke, Magnar (2006) Domain-Focal Terms and the Lexical Delimitation of Subdomains. In: Picht, Heribert (ed.). *Modern Approaches to Terminological Theories and Applications*. Bern: Peter Lang. 355-374.

Gibbons, John (1994) *Language and the Law*. London: Longman.

Hoffmann, Lothar (1985) *Kommunikationsmittel Fachsprache. Eine Einführung*. Berlin: Akademie-Verlag.

Kotler, Philip (2003) *Marketing Management.* New Jersey: Pearson Education, Inc.

Kotler, Philip (2004) *Markedsføringsledelse.* Oslo: Gyldendal Norsk Forlag AS.

Kristiansen, Marita (2004) *The Multi-Disciplinary Nature of the Social Sciences. Investigating Disciplinary Autonomy in Organisational Behaviour by means of Terminological Analysis.* Bergen: University of Bergen/NHH, Dr.art thesis.

Kristiansen, Marita (2005) Disciplinary autonomy and concept relations in electronic knowledge bases. A theoretical approach to KB-N – a knowledge base for economic-administrative domains. *SYNAPS – Fagspråk. Kommunikasjon. Kulturkunnskap,* 17(2005). 1-7.

Love, Stacey (2000) Benchmarking the Performance of Two Automated Term-Extraction Systems: LOGOS and ATAO. URL http://www.olst.umontreal.ca/pdf/memoirelove.pdf <16.06.08>

Norman, Guy J. (2003) Consistent naming in scientific writing: sound advice or shibboleth. *English for Specific Purposes*. 113-130.

Øvsthus, K[ari] /Innselset, K[ai] /Brekke, M[agnar] /Kristiansen, M[arita] (2005b) Developing Automatic Term Extraction. Automatic Domain Specific Term Extraction for Norwegian. In: Madsen, Bodil Nistrup / Thomsen, Hanne Erdmann (eds.). *Terminology and Content Development. TKE 2005. 7th International Conference on Terminology and Knowledge Engineering*. [s.l.]: Litera. 337-348.

Picht, Heribert (1996). Fachkommunikation – Fachsprache. In: Budin, Gerhard (ed.) *Multilingualism in Specialist Communication. Proceedings of the 10th European LSP Symposium, Vienna, 29. Aug.-1. Sept. 1995*. Vienna: International Network for Terminology. 27-45.

Picht, Heribert (2007) *Introduktion til terminologisk metode og dens grundbegreber.* URL http://gandalf.aksis.uib.no/nordterm/presentasjoner/Kursus_H_Picht_2007.ppt <03.06.08>

Rogers, Margaret (2007) Lexical chains in technical translation. In: Bassey, Edem Antia (ed.) *Indeterminacy in terminology and LSP*. Amsterdam/Philadelphia: John Benjamins. 15-34.

Sager, Juan /Dungworth, David /McDonald, Peter P. (1980) *English special languages. Principles and practice in science and technology.* Wiesbaden: Brandstetter.

Stubbs, Michael (1996) *Text and Corpus Analysis. Computer-assisted Studies of Language and Culture*. Oxford: Blackwell.

The Norwegian newspaper corpus, Aksis/Unifob AS, URL http://avis.uib.no/english.page <03.06.08>