

Skimming for Paragraphs

Relating of Concepts

Postgraduate paper in
Computer Science

Ole Torp Lassen
DIKU, Spring 2006

Advisors:
Neil D. Jones
Peter R. Skadhauge

Skimming for Paragraphs.

Introduction	2
1 Partitions as scopes of unambiguity.....	3
2 Finding and comparing partitions.	4
2.1 Realizing the subtasks.	4
2.2 Choosing starting point and direction. ...	7
2.3 Refining the approach.....	18
3 Informal results and conclusions.....	22
3.1 Conclusion.....	23
Appendix	25
A-1 References	25
A-2 The experimental corpus, CivIII.....	25

Introduction

This postgraduate written assignment is to be seen as a continuation of the work done in my thesis paper (Lassen, 2005). While that paper dealt with the development of a prototypical system for recognising semantic context of informative English text, this paper deals with a possible application of such a system, namely the structured semantic partitioning of texts.

I will in this assignment try to provide a tentative answer to the following question:

“Can skimming as presented in (Lassen, 2005) possibly be applied in order to arrive at a semantically well motivated partitioning of informative English text? “

I will analyse the different aspects of the problem and if possible provide an algorithm for its solution. While I will discuss the quality of the result, I will not attempt to prove it, as this would make the paper grow out of proportions. I will however touch on how such “proof” may be found through intensive experimenting.

Therefore, rather than a presentation of final project results, this paper represents a research proposal providing a thorough tentative analysis of the problem, as also was the case for (Lassen, 2005).

1 Partitions as scopes of unambiguity.

In my thesis, (Lassen, 2005), I described an experimental method for assigning contextual representations to portions of natural language text. The method, called *skimming*, involved representing each data text as the sequence of its nouns in orthographic form. The algorithm attempts to assign meanings to as many of the nouns as possible through analysis of semantic coherence as expressed by semantic relations, especially hyperonymy, between the possible interpretations of the nouns. Furthermore, the following restrictions were placed on the possible solutions:

- a) Any meaning assigned to a noun in the sequence must be semantically related to the assigned meaning of at least one other noun in the sequence.
- b) No word is allowed more than one *assigned* meaning within the same *paragraph* of the text. (While a word may have several possible meanings, at most one will be allowed in any given interpretation).

The result of this exercise is a set of *lexemes*, i.e.: a set of pairings of words to meanings, along with a set of semantic relationships connecting those lexemes. I showed how this set of lexemes could be seen as a sketchy representation of the semantic context of the text.

The restriction in b) above, touches on an obviously important issue in natural language semantics: many if not all words of natural language are ambiguous, some more often than others. It is, however, reasonable to assume a **scope of unambiguity**, i.e.: a certain scope of linguistic interactivity, within which words are intended as unambiguous and where they can safely be regarded as such.

Intuitively this notion is closely related to the notion of semantic context and obviously one of the reasons why NL-systems tend to restrict themselves to analyse data from clearly defined and distinct semantic domains. It seems reasonable to assume that the scope of unambiguity extends conceptually to what is conceived as one context (even though it may comprise an entire cluster of semantic contexts, as long as they do not contradict each other). In these terms, it must be a key ability of a generally applicable NL-system to be able to analyse linguistic data “one scope at a time”.

In (Lassen, 2005) the typographical paragraphs of the text itself were chosen as reasonably good indications to the contextual boundaries of the text. Thus each typographical paragraph was treated as one scope of unambiguity – one unambiguous cluster of contexts. While there is good reason behind assigning importance to the typographical paragraphs of a text like this, it does involve a couple of important problems:

- Firstly, the use of paragraphs in a text is a question of style rather than one of convention. That is, we can't rely on a consistent use and meaning of typographical paragraphs.
- Secondly, not all formats of language support the use of typographical paragraphs, dialog transcriptions are obvious examples, but also many linguistically prepared corpora of tagged text does not employ reference to paragraph boundaries.

In this paper I want to explore the possibility of applying skimming in order to partition a text into semantically grounded portions without reference to more or less coincidental typographical paragraphs that may or may not be present in the original text.

2 Finding and comparing partitions.

Basically the present experiment involves elimination of any and all typographical paragraph markers in a copy of the small CIVIII-corpus that I used in "Skimming for Context". This leaves us all the nouns of the entire corpus (i.e.: from paragraphs 0 through 8) in sequence.

The task is to devise an algorithm to break this long sequence up in semantically well-motivated partitions, each corresponding to one scope of unambiguity - one cluster of contexts that do not contradict each other.

The newfound partitions can then be compared to various control-partitions and the relative deviation could be regarded as a measure of success. The simplest such set of control partitions is of course the original partitioning of the author, namely the set of typographical paragraphs that were eliminated from the original text.¹

The rest of this chapter will discuss how this goal might be achieved.

2.1 Realizing the subtasks.

Since I want to apply skimming in a clever way in order to divide a text into meaningful paragraphs, I should summarize how skimming works.

- The skimming algorithm requires a sequence of nouns. It regards that sequence as one scope of unambiguity and offers the "best" partial interpretation of the nouns under those circumstances.

ⁱ A more reliable control set should of course involve a (probably large) number of test-persons and a generalisation of their preferred partitioning of the original text. This being said, the original set of paragraphs should suffice as a rough measure of the performance of the basic idea. As was the case in the main thesis, I want to keep things as simple as possible and thus I refrain from large, time consuming subtasks where useable alternatives are readily available.

- The “best” partial interpretation is decided, by comparing the number of lexemes and instances involved in competing interpretations.
- A partial interpretation is represented as a set of lexemes ranging over the nouns of the sequence and a set of semantic relationships between those lexemes.
- It is seldom the case that all the nouns in a sequence are assigned a meaning by the algorithm, hence the term *partial interpretation*.

As mentioned earlier, the typographical paragraphs of the original corpus text were themselves regarded as semantic units that could be grouped together as shown in **Figure 1**. Here is sketched a sequence involving nine typographical paragraphs, p0-p8. The boxes indicates how the paragraphs could be organized in several different ways: each in isolation, in various groupings or all in one. If each typographical paragraph is indeed semantically well motivated it makes good sense to use them as possible partition lines, e.g.: if each paragraph is intended (and can be treated) as one scope of unambiguity. On the other hand, we have no guarantee that the typographical paragraphs placed by the author of the original text can be treated this rigorously. The

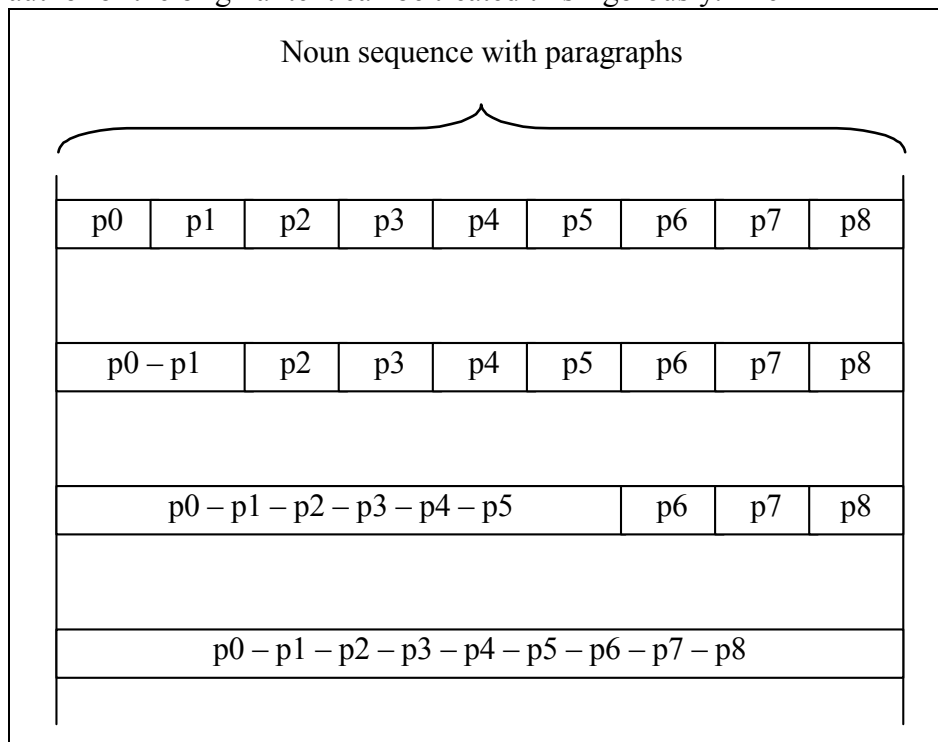


Figure 1: The sequence of nouns divided into portions according to the typographical paragraphs of the original text. Regarding the typographical paragraphs of the original corpus text as semantic units on their own provides for semantic partitions that are easily experimented with.

typographical layout of a text remains a question of style as already mentioned. Therefore it is of obvious interest to be able to recognize scopes of unambiguity without reference to typography. This ability would allow for well motivated semantic units to be available in a more reliable way, even in the absence of typographical paragraphs in the original text. This means that, at least for now, we will have to eliminate all references to typography in the source text, as sketched in **Figure 2**. Having disposed of the original paragraphs of the text we can begin to explore how to partition the text in a way, that reflects the semantic content of the text in a consistent and reliable manner.

Given a sequence of two or more nouns and regarding it as one scope of unambiguity, the skimming algorithm assigns a contextual representation to the sequence. This intuitively indicates that the noun sequence can be broken up into two or more sub-sequences that can then be skimmed separately in order to access the information contained in their respective contextual representations. Generally, a different partitioning of the noun sequence will produce a different set of contextual representations and thus provides grounds for objective

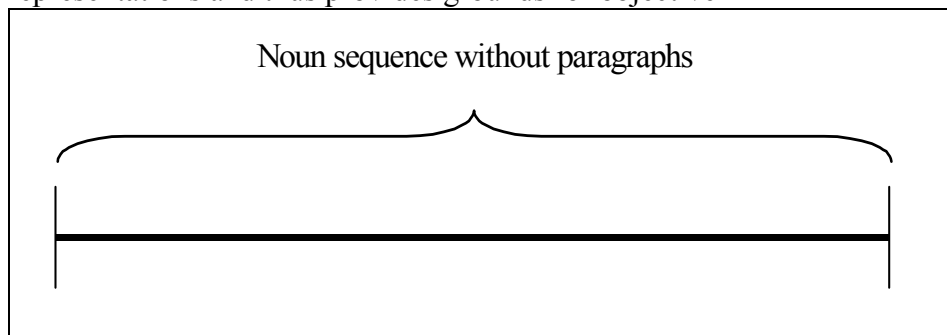


Figure 2: All typographical information has been eliminated.

comparison of partitions. If we can find one well motivated partition in the noun sequence, finding the rest is a mere question of repetition as necessary. So, what we need to do is essentially the following three subtasks :

- 1) **Decide on starting set of partitions.** Once we have decided on a starting partition set, we can start skimming it.
- 2) **Decide on transition between partition sets.** Once we have decided on how to go from one partition set to the next, we can start skimming the alternative partitions.
- 3) **Decide on a measure of comparison for partition sets.** Once we have decided on how to compare partitions and partition sets we can decide when to look for a new transition to improve the current partition and when to stop because no further improvement is possible.

2.2 Choosing starting point and direction.

Since the skimming algorithm requires a sequence, with which to start its search, it is necessary to decide on how to begin. Of course, deciding on how to begin also places restrictions on how to proceed. There are several ways one can go about this task but basically, alternative approaches must divide themselves between what I will call the *restrictive* and *expansive* approaches. These paradigms will be the focus of discussion in the following subsection.

2.1.1 The restrictive approach

- start with the largest possible partition and gradually restrict its boundaries until done.

The sequence S is assumed to be representative of a coherent source text, rather than a selection of random nouns and as such it makes sense to regard the entire sequence as one scope of unambiguity. The question at hand is whether it makes better sense to regard the sequence as several, smaller such scopes of unambiguity in sequence, like beads on a string.

The restrictive approach to decomposition of the sequence S is sketched in **Figure 3**. It starts by skimming the entire sequence and record the resulting contextual representation. The intuitive next step would be to gradually restrict the boundaries of the starting partition by skimming smaller and smaller partitions and comparing the respective results until one scope of unambiguity has been reached. It should be apparent however, that there are a couple of serious problems with this approach. Firstly, starting the search for a comparably small portion of S by treating all of it, obviously involves a lot of unnecessary work.

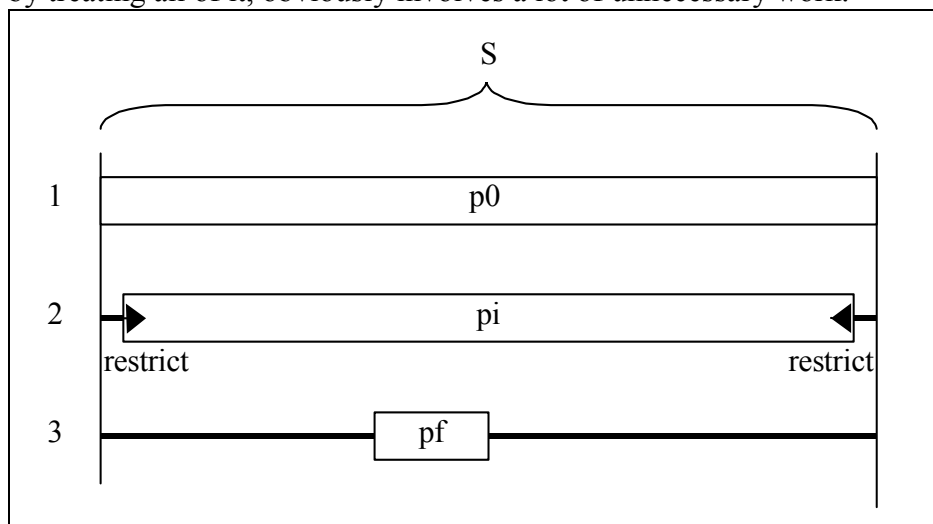


Figure 3: Intuitively the restrictive approach begins by skimming, the largest possible portion of the sequence, namely the entire sequence S , as the starting partition. The next step could be to restrict the boundaries of the partition and skim the resulting partitions until one partition has been found that holds only one scope of unambiguity.

Assuming that it is possible to find a well motivated paragraph this way, still only one is found. That means that to find more, the process will have to be repeated by skimming the sequences before and after the newly found partition in their entirety (remember that both of these have already been involved in a skimming process, finding the partition between them). Having found two new well motivated paragraphs this way, the process will still have to be repeated until all of S has been marked as belonging to one such paragraph. On each such repetition, the largest possible start partition is chosen, resulting in potentially large portions of S being skimmed multiple times.

Secondly, it may well prove a very difficult task to decide when the current partition does indeed involve only one scope of unambiguity. This becomes completely clear when regarding the case where the entire sequence S is the only scope of unambiguity possible - where there is only one partition in the sequence, namely the sequence of S itself - the one that we actually start out with. Even in this case we still have to repeatedly chop off nouns two by two, meticulously skimming and comparing the intermediate results, all the way down to a partition of two nouns, just to make sure that there isn't more than one partition involved in S.

Suppose the problems are associated with the declared focus on the extent of partitions. What if we instead focus on the dividing line between possible partitions of S. Instead of explicitly restricting the starting partition, p0, we could try by inserting a dividing line somewhere in the sequence, just to see if it might make sense to split the sequence up and interpret the sub-sequences as separate scopes of

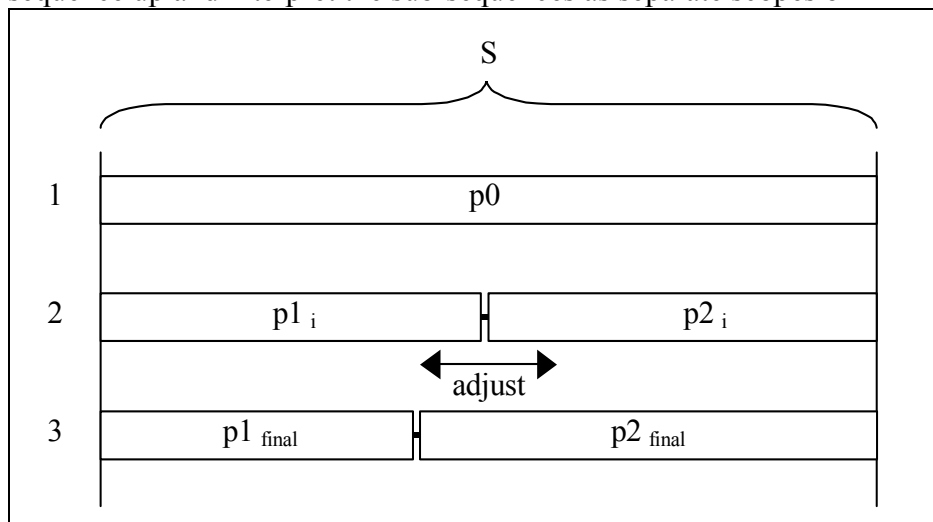


Figure 4: Splitting S into two smaller sequences p1 and p2, represents an attempt to see if the interpretation of S would benefit from the introduction of an additional scope of unambiguity. Skimming p0, p1 and p2 and comparing the results should reveal such a benefit. If there was a benefit the relative proportions of the respective partitions may be adjusted in a similar way.

unambiguity as sketched in **Figure 4**.

Well, no it doesn't. Indeed, in the case, where that is only one partition involved in S, we still have to try all possible boundaries to make sure. Furthermore, the problem of starting each repetition by the worst possible workload still remains, and has indeed worsened considerably, since all the nouns of S will now be skimmed every time the process is repeated, instead of just a smaller and smaller portion of them. While there may be slight differences between the operational behaviour and complexity of these two procedures, their basic problems are intrinsic to the restrictive approach, that both are variations of.

In hindsight, the restrictive approach is probably best suited for finding a comparably small number of comparably large objects. It is not at all suited for finding many, potentially small objects.

2.1.2 The Expansive approach

- start with the smallest possible partition and gradually expand its boundaries until done.

By now it seems clear that starting with the smallest possible partition will mean finding the partitions from the inside and out and reserve the worst possible workload for the worst possible case, so to speak. To realise the size of such a partition, recall that the skimming

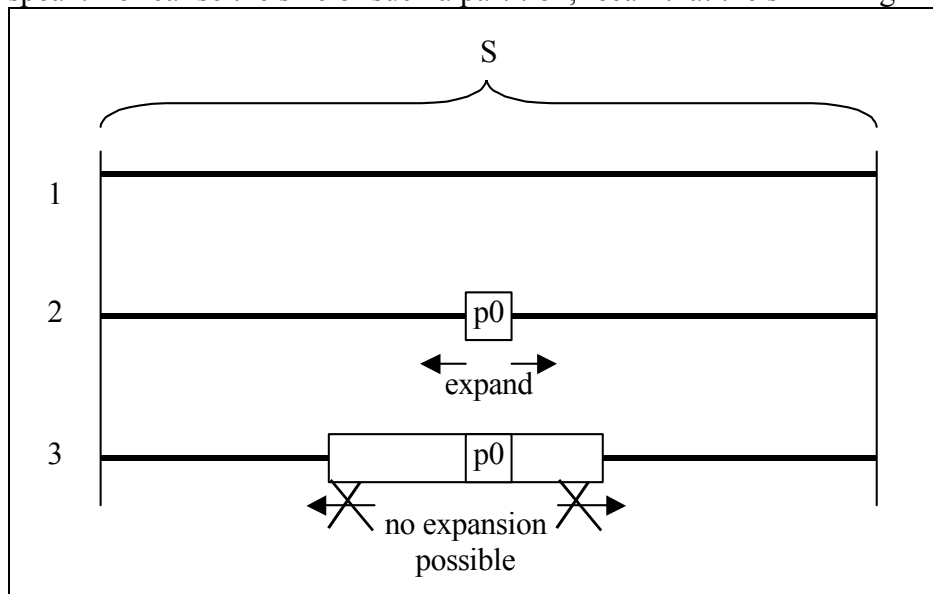


Figure 5: The expansive approach begins by skimming the smallest possible start partition. Since the skimming algorithm requires a sequence of at least two nouns to form a nontrivial contextual representation, the smallest possible start partition could be any pair of consecutive nouns in the sequence, for instance p0 in the diagram. Intuitively, the expansive approach must proceed by expanding p0, skim the expansion and compare the respective contextual representations. This process must continue as long as the expansion doesn't result in a generalisation of the contextual representation.

procedure depends on semantic relationships between possible interpretations of nouns in the sequence. Because any relationship requires at least two participants, the smallest nontrivial partition of nouns to be skimmed is a sequence of two nouns.

However, some redundant work is still done. Suppose for instance that p_0 is the starting pair of nouns that expands to a partition of ten nouns in total, four on each side of p_0 . The expansive approach starts by skimming p_0 , i.e.: a sequence of two nouns. Let's say that we now expand p_0 in both directions to a new sequence and comparing the respective results to the result of skimming p_0 could then serve as an indication of the potential benefit of partitioning to the interpretation of the sequence.

So now that we only have to decide the relative soundness of the position a single partition boundary, does this offer relaxation of the workload? So, as sketched in **Figure 5**, the expansive approach begins by skimming a sub-sequence of S of length 2ⁱⁱ. The sub-sequence will then be expanded in either or both directions as long as the result of skimming the expanded partition is at least as "good" as the previous result. The moment it can be discerned that expanding the partition in either direction will result in a generalisation of the contextual representation, we are done. Generalisation of the contextual representation can be seen as an indication that a boundary between scopes of unambiguity has been breached.

Since this approach attempts to find partitions from the inside, the only case where the expanding/skimming cycle involves the entire sequence of S , is when there is only one meaningful partition in S , namely S itself. Clearly, the portions of S that are treated by this approach will start out small and gradually grow up until the point where the boundaries to the neighbouring partitions are met. sequence, p_1 , of four nouns, namely p_0 and the noun on either side of p_0 . Skimming p_1 now means skimming two new nouns and two that have already, skimmed once before. This will repeat itself until partition p_5 of length 12, that represents a generalisation. At this point, it is necessary to see what new noun in p_5 caused the generalisation, the first, the last or both. This means that two sequences each of eleven nouns have to be skimmed again, even though they have all been skimmed before as part of (an) earlier expansion(s). Basically, to find a partition of m nouns, we have to expand the start partition from 2 to $m+2$ nouns. To do that, each noun will on average be involved in $(m+4)/4 + 2$ skimming operations. The necessity of this double work stems from the fact that the experimental skimming prototype still works statically rather than

ⁱⁱ This way, it is actually assumed that the two nouns belong to the same scope of unambiguity even though this might not be the case. In the special case that the start nouns belong to different scopes or partitions the result has to be regarded with special care. It remains to be seen if a partition resulting from such a false premise can be recognised to be anomalous in some way. Generally there is a comparably high probability that any two consecutive nouns does indeed belong to the scope of unambiguity. I will return to this matter later in this paper.

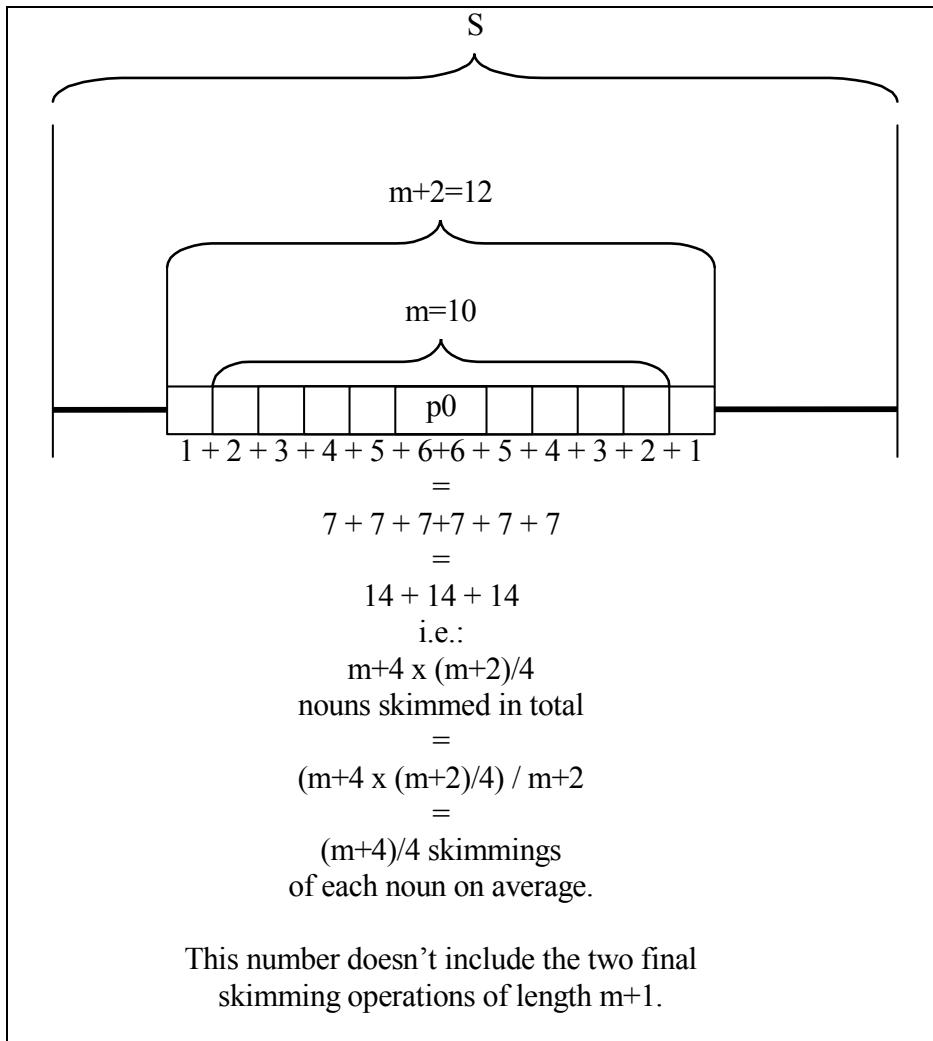


Figure 6: Each noun involved in finding a partition of 10 is skimmed on average almost 5.5 times by 8 separate skimming operations.

dynamically. If the skimming algorithm had been refined to work dynamically, the expansive approach would only have to skim each noun once. Incidentally, to compare a sequence to the result of expanding it, it is necessary to treat both sequences, as all together different sequences and thus, to skim the overlap somewhat redundantly.

But, rather than a problem intrinsic to the expansive approach, this redundancy is a consequence of the current state of the experimental skimming algorithm and cannot, at least for the time being, be avoided. Furthermore, the expansive approach actually does the better job minimizing the consequences by keeping the sequences to be skimmed as short as possible. Finally, as sketched in **Figure 6**, the increase in complexity clearly remains polynomially proportional to the size of S (i.e.: the redundancy doesn't cause the problem to become intractable).

Consequently, the conclusion to this discussion must be, that even though both approaches involve redundancy to a certain extent, the

expansive approach is the best suited to find partitions. With reference to the subtasks listing of section 2.1, this means that we have decided that :

- 1) The start partition will be a sub-sequence of S of length 2.
- 2) The stepwise transition from partition to partition will be that of expanding the sequence.

Before we can elaborate further on this general paradigm, we must decide on how to compare the competing partitions.

2.3 Comparing partitions.

Having decided how to begin and proceed we now need to decide when to stop. To make this decision, we must first realize the kind of changes that expansion can impose on the contextual representation. As already noted, the contextual representation of a sequence, is in essence the graph resulting from skimming that particular sequence, so let us try to expand an example partition in various ways and discuss the respective changes in context.

Suppose we have a sequence, S, of nouns that we wish to partition. We start by choosing some arbitrary start-partition, p0, of S. Suppose that p0 consists of the nouns *civilization* and *society*. Now, assuming that the words in p0 go together in the same scope of unambiguity, we skim p0 and get as a result the best interpreting graph of p0, I(p0), that we regard as the contextual representation of the partition. We see that, the current context involves two lexemes (society,s10), (civilization,s1) and a relational edge of hyponymy between the two. Furthermore we see that the interpretation has a score of 2 and that it is complete, i.e.: it assigns meaning to all the nouns of p0. This is illustrated in **Figure 7**. A glossary for this example can be found in **Figure 11**.

We can now expand p0 in different contexts to see how they result in different contextual representations.

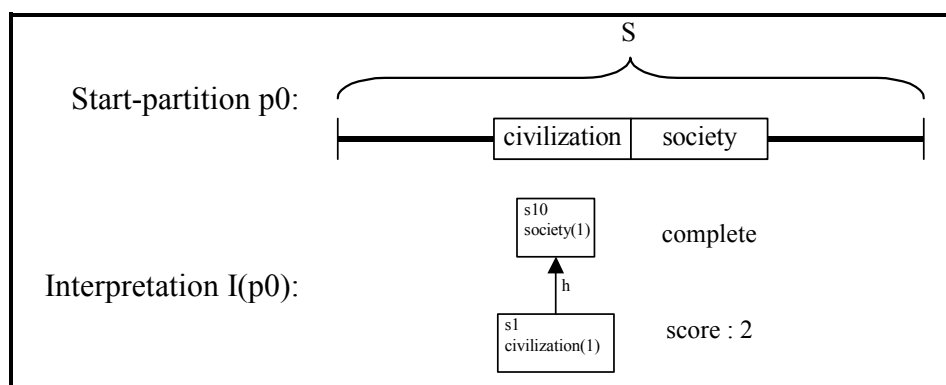


Figure 7: Example start-partition and its contextual representation

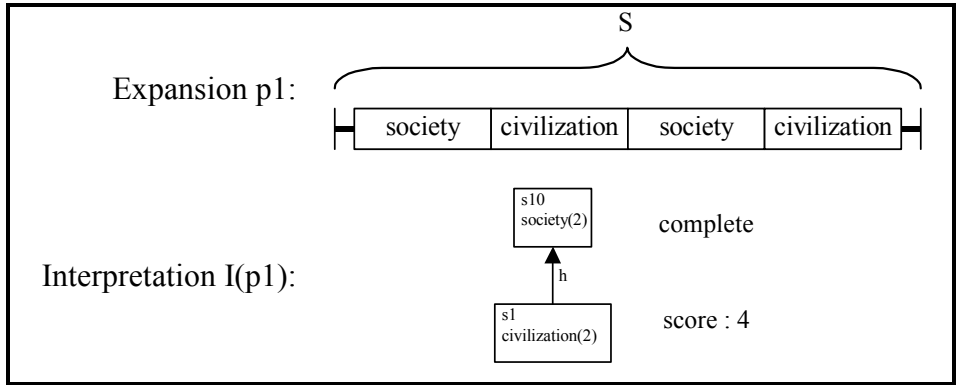


Figure 8: This expansion *confirms* p0.

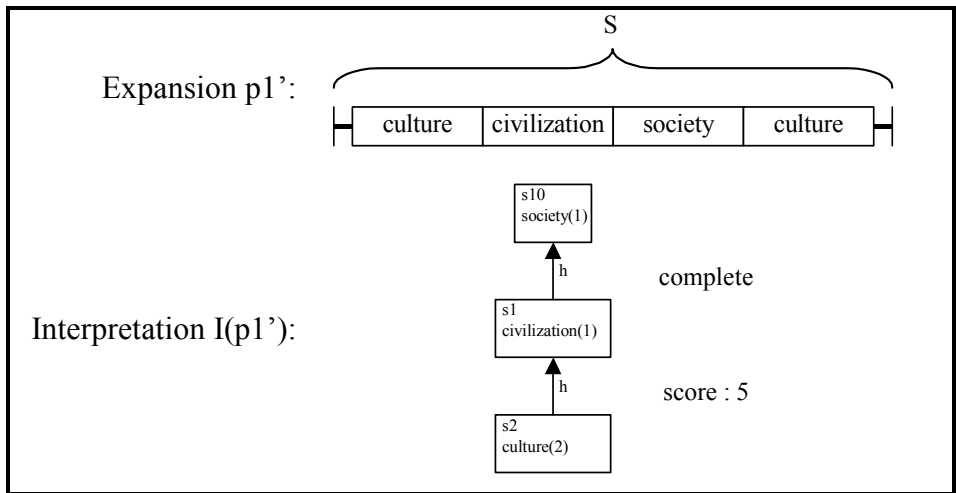


Figure 9: This expansion *refines* p0.

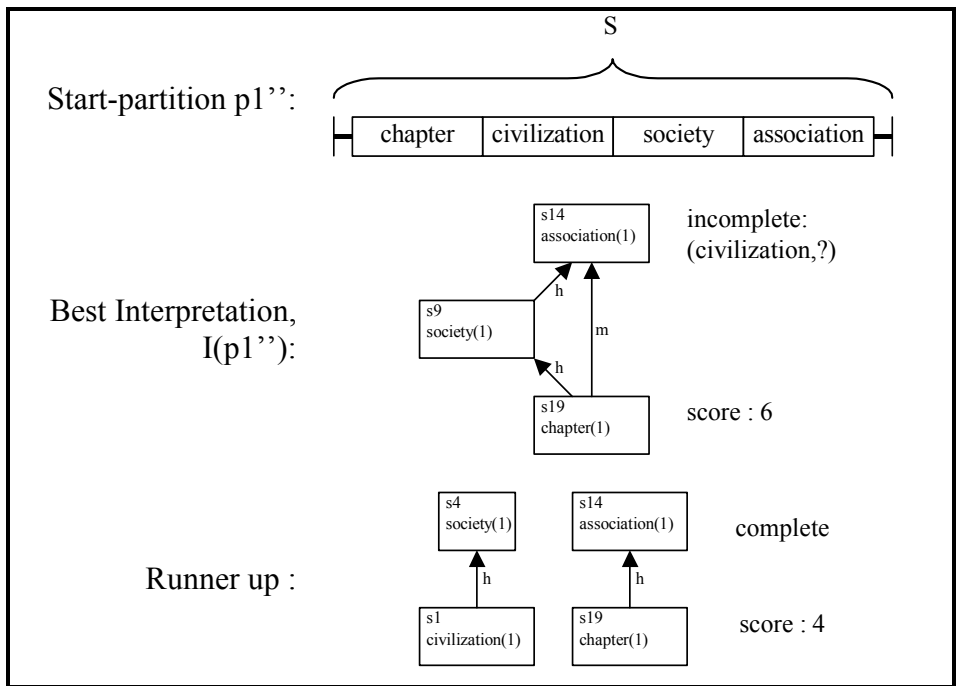


Figure 10: This expansion is a *generalisation* of p0, since the established context is discontinued. The conservative interpretation scores to low.

Glossary for the nouns:

civilization: s1, culture: s2-s5, society: s6-s10, association: s11-s16, chapter: s17-s21

handle explanation

- s1** - a society in an advanced state of development.
- s2** - a particular civilization at a particular stage.
- s3** - all the knowledge and values shared by a society.
- s4** - the tastes in art and manners that are favoured by a social group.
- s5** - (biology) the growing of microorganisms in a nutrient medium - such as gelatin or agar.; "the culture of cells in a Petri dish".
- s6** - the raising of plants or animals: "the culture of oysters".
- s7** - the state of being with someone; "he missed their company"; "he enjoyed the society of his friends".
- s8** - the fashionable elite.
- s9** - a formal association of people with similar interests; "he joined a golf club"; "they formed a small lunch society"; "men from the fraternal order will staff the soup kitchen today".
- s10** - an extended social group having a distinctive cultural and economic organization.
- s11** - the state of being connected together as in memory or imagination; "his association of his father with being beaten was too strong to break".
- s12** - a social or business relationship: "a valuable financial affiliation"; "he was sorry he had to sever his ties with other members of the team"; "many close associations with England".
- s13** - any process of combination - in solution. that depend on relatively weak chemical bonding.
- s14** - a formal organization of people; "he joined the Modern Language Association".
- s15** - the process of bringing ideas or events together in memory or imagination; "conditioning is a form of learning by association".
- s16** - the act of consorting with or joining with others; "you cannot be convicted of criminal guilt by association".
- s17** - a distinct period in history or in a person's life; "the industrial revolution opened a new chapter in British history"; "the divorce was an ugly chapter in their relationship".
- s18** - an ecclesiastical assembly of the monks in a monastery or even of the canons of a church.
- s19** - a local branch of some fraternity or association; "he joined the Atlanta chapter".
- s20** - a series of related events forming an episode; "a chapter of disasters".
- s21** - a subdivision of a written work; usually numbered and titled; "he read a chapter every night before falling asleep".

Figure 11: Glossary for the five example nouns.

First, assume that the partition p_1 in **Figure 8** is the result of expanding p_0 one word in both directions. We see that the words in p_1 duplicates the words in p_0 and consequently, the structure of the interpreting graph remains the same. We clearly would want the system to recognize p_1 as a continuation and thus an acceptable expansion of p_0 . While there is a score associated with the interpretations, this score was computed in order to compare alternative interpretations of the same sequence or partitions rather than to compare different partitions. Instead we must turn to the contexts represented by the respective graphs. To this end, it should be clear that it suffices to keep track of the lexemes involved in the respective interpretation. The set of lexemes involved in $I(p_0)$ is as follows :

$$L_0 = \{(s1, civilization), (s10, society)\},$$

while the set involved in $I(p_1)$ is:

$$L1=L0=\{(s1,civilization),(s10,society)\}.$$

Now, instead of $p1$, assume the partition $p1'$ in **Figure 9** as the result of expanding $p0$. Again we expand the original partition, $p0$, one word in both directions. This time the result is the addition of the word *culture* to each end of $p0$. Where $I(p0)$ and $I(p1)$ were identical, $I(p1')$ has an added culture-vertex to the component of *society* and *civilization*. We see that :

$$L1' = \{(s1,civilization),(s2,cuklture),(s10,society)\}$$

and

$$L0 \subseteq L1'$$

Since all lexemes in $I(p0)$ persist in $I(p1')$, we would want to accept $p1'$ as a proper expansion of $p0$, just as well as $p1$.

Finally, regard the example in **Figure 10**. Here we expand $p0$ with the words *chapter* and *association* resulting in the partition p'' . These words have meanings that relate to each other and also to one particular meaning of *society*. Since that meaning of *society* is different from the one involved in $I(p0)$ we should see a discontinuation of that partition in contrast to the previous examples. When regarding the preferred interpretation of p'' , we see that those three words can be interpreted in very strong relation to each other as one coherent component, while *civilization* is left un-interpreted as a consequence, since it has no related meaning in the respective context. The new lexeme set looks like this :

$$L''=\{(s14,association),(s19,chapter),(s9,society)\}$$

i.e.: $L0 \not\subseteq L''$, in fact the two sets are completely distinct.

Clearly, this is the kind of behaviour that should raise the alarm that a boundary between two separate scopes of unambiguity has been encountered. The proper response should be to undo the trespassing expansion and accept the previous partition as non-expandable. To be orderly, I must mention that a weaker alternative interpretation, represented as “The runner-up” in **Figure 10**, was in fact considered for best interpretation of p'' . This interpretation continues the context of $p0$ and introduces a new *chapter-association* component instead of reinterpreting *society*. $I(p'')$ comes out the winner, however, because of the strong coherency indicated by the extra relational edge.ⁱⁱⁱ

ⁱⁱⁱ While the alternative interpretation of p'' continues the context of $p0$ without contradictions, the skimming algorithm decided that $I(p'')$ is the better interpretation of the sequence. Whether this is reasonable or not is an issue relating to how to refine the skimming algorithm and will not be discussed here. The example is a general one that illustrates how to possibly recognize contextual changes of gradually expanding

- I. The contextual representation of the original partition does not change from the proposed expansion. This can happen in two cases :
 - a) The encountered word has no interpretation that relates to any possible context. While it doesn't promote the established context it doesn't contradict it either. Being neutral in the strife between competing contexts, I will refer to this as a **continuation**.
 - b) The encountered word has an earlier occurrence in S, and it has already been interpreted and represented in the established context. While it doesn't add anything new to the context it strengthens the established context through the repetition. Presenting a stronger argument in favour of the established context than continuation I will refer to this as a **confirmation (Figure 8)**.
- II. New lexemes are introduced to the contextual representation while original ones remain. This is the strongest possible argument in favour of the established context and is what I will refer to as a **refinement** of it (see **Figure 9**).
- III. While the actual number of lexemes may grow, some original lexemes may "fall out" or get re-interpreted as a result of the expansion. This is the what I refer to as **generalisation** of the contextual representation (**Figure 10**) and clearly argues that expansion as gone to far.

Figure 12: Distinctions of the possible consequences of expansion with respect to the established context

Summing up, it should be clear from this small example, that several things can happen when comparing a partition and its proposed expansion. In particular, by monitoring the lexemes involved in the respective partitions it the following distinctions can be made:

Both I and II above should, in this respect, vouch for the proposed expansion. In both cases we see that all lexemes of the established context persist through the proposed expansion, i.e.: L_i is a subset of L_{i+1} . The Generalisation in III, on the other hand, indicates that a scope boundary has been crossed; it should prohibit the proposed expansion and instead accept the previous partition as un-expandable and end the procedure. A case of generalisation is recognised simply by finding that L_i is not a subset of L_{i+1} .

In terms of argumentative strength continuation is of course somewhat weaker than both confirmation and refinement. In terms of relative frequency of occurrence, I expect instances that cause

partitions in a consistent way, and here we can simply pretend that the truth of the skimmer is absolute.

continuation or confirmation to outnumber those that refine or generalise the established context.

We can now formalize the decision of when to continue the expansion process and when to stop in terms of the subset relation as follows: with partition p_i and its potential expansion p_{i+1} do comparison on the lexeme sets, L_i and L_{i+1} , of their respective interpretations.

If $L_i \subseteq L_{i+1}$, accept p_{i+1} and continue
 else return p_i

This way the comparison of partitions can be reduced to a comparison of sets of lexemes, neatly reflecting the underlying conceptualisation of **scope of unambiguity** and its relation to that of **semantic context**.

With reference to the subtasks listing of section 2.1, we have now completed the third and last of the preliminary subtasks :

The start partition will be a sub-sequence of S of length 2.

The stepwise transition from partition to partition will be that of expanding the current partition. Expansion will continue as long as all emerging lexemes persist. Stop once expanding the current partition would cause a generalisation of the current set of lexemes or when the boundaries of S are encountered.

Having decided on an non-deterministic approach to partitioning I can now begin the discussion of how to achieve determinism in the matter.

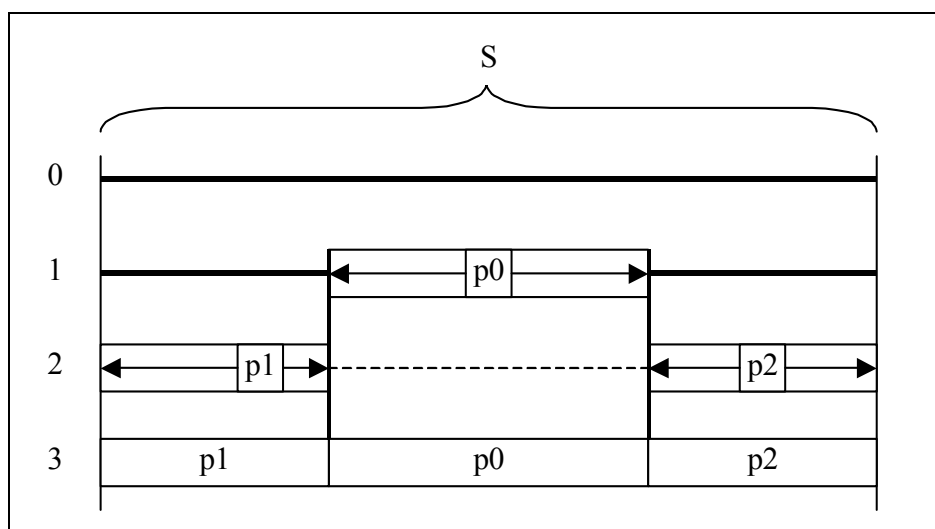


Figure 13: Intuitively, finding the first partition, p_0 , and then repeating the process on the remaining sub-sequences of S, should provide determinism. But are boundaries between contexts really this clear cut ?

2.3 Refining the approach.

It might seem that achieving determinism in this case is simply a case of applying the non-deterministic algorithm recursively to the portions of S that have not already been assigned to a partition until all of S is covered. As shown in

Figure 13, this makes for a series of clearly distinguishable partitions to represent the semantic organisation of the original text. If indeed, the respective boundaries found by the algorithm were absolutely accurate with respect to the respective contextual foci of the original text, this would be sufficient. However, the boundaries can not be regarded as accurate for several reasons. Most importantly, because the partitions are found through expansion from particular starting-points, the resulting scopes of unambiguity each represent the most extreme such expansion possible. More precisely, each boundary marks the widest expansion from its respective starting-point in a particular direction that can be made without generalisation of the established context.

That is to say that at the very latest, there must be a boundary at this point with respect to that particular starting point and direction. In terms of the distinct consequences introduced in **Figure 12**, the closest we can reasonably get to the actual position of the boundary is to claim that it must be somewhere in span of continuations between the latest refinement/confirmation of the established context and the earliest generalisation of it.

This means that p0 in **Figure 13** quite likely overlaps p1 and p2 to some extent and that it is not “fair” to let the bounds of one partition restrict the possible expansion of its neighbours, just because it

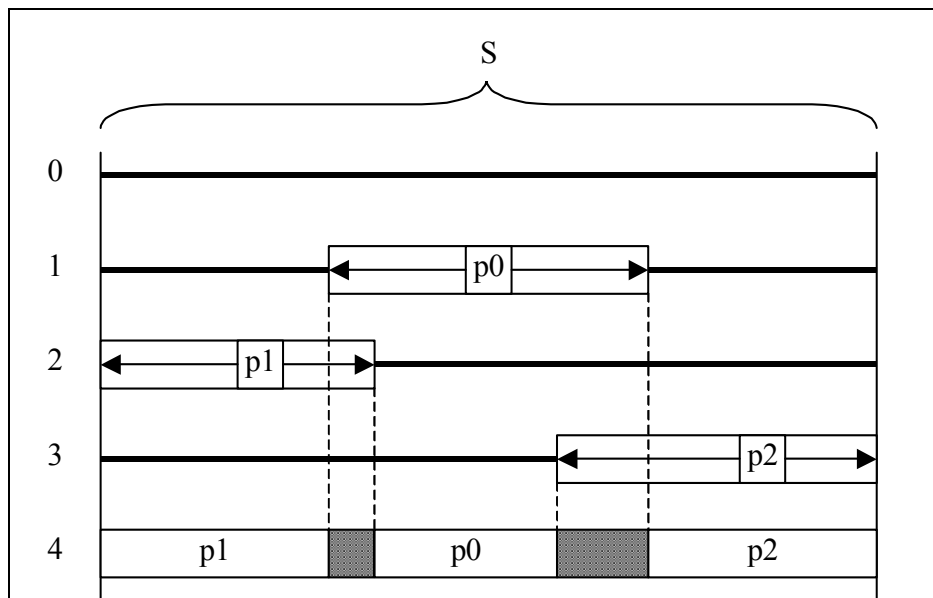


Figure 14: When each algorithmic cycle receives its own uncluttered version of S, scopes can expand freely to their fullest extent . Perhaps a better representation of the boundaries between scopes is emerging.

coincidentally was expanded first. Instead each scope should be allowed to expand unrestricted, as in **Figure 14**, i.e., with no regards to what other scopes has or has not already been recognised. It should be clear that allowing scopes of unambiguity to expand independently of each other provides for the clearest picture of how they may interact. In addition, it is entirely possible that the overlap between scopes is the best bet as to the actual position of the contextual shift between the scopes occurs, and may indeed further restrict the span of continuations between refinement/ confirmation and generalisation, that I mentioned earlier. That the overlap between p1 and p0 in Figure 14, for instance, holds the actual boundary between the respective scopes seems obvious since the right extreme of p1 is latest possible position for p1's boundary to p0, the left extreme of p0 is the latest possible position for its boundary to p1 – ergo, their common boundary must be in between the two extremes, including both of them.

As a further consequence of this potential overlap between neighbouring scopes, we must try each and every start partition to make ensure recognition of all scopes of unambiguity in S. Even though this certainly will find each scope several times, we can not simply restrict ourselves to try only start partitions that have not already been involved in other scopes. While this sounds unnecessarily cumbersome it is the only way we can monitor how the respective scopes of unambiguity behave and interact. These and further experiments may provide clues on how to better partition linguistic data according to semantic content, but for now we must be consequent and make sure we see it all. Therefore, I will begin with the earliest possible start partition of S, namely the sequence consisting of the first and second nouns of S. After expanding this partition as far as possible and recording the result, I will regard the start partition one noun to the right of the first one, to the second and third nouns of S, and expand it. This, I will continue all the way until the n-1'th and n'th noun of S has been tried as start partition.

2.4 Summing up and representing the results.

The conceptual analysis of the problem is drawing to an end and an experimental algorithm for finding scopes of unambiguity in a natural language text can be formulated as in **Figure 15**.

This iterative application of the skimming prototype is easily written and will for each iteration produce a triple like this :

$((X,Y),L,R)$, where

- (X,Y) is the starting partition of the iteration represented by the left and right extremes in S.
- L and R are the leftmost and rightmost expansions from that starting partition as described.

Since all possible pairs (X,Y) of consecutive nouns in S, will be tried as starting partitions in this experiment, I will end up with n-1

overlapping scopes of unambiguity in S to compare. As a theoretic example and illustration of this, Figure 16 shows a sketch of how different contextual behaviours might be distinguished by examining how the scopes of unambiguity distribute themselves over the respective data sequence. First imagine a sequence of 10 nouns where the scopes of

- 1: Starting with the pair of first and second noun in S, regard them as part of the same scope of unambiguity and accept this pair as **starting partition** and **current partition**. Skim **current partition** and remember the lexemes of the best interpretation as **current lexemes**.

↔ ?
- 2: If possible, expand **current partition** in both directions, skim the **new partition** and regard the **new lexemes**,
else goto 4.
- 3: If **current lexemes** \subseteq **new lexemes** then
 remember **new lexemes** as **current lexemes**,
 remember **new partition** as **current partition**
 and goto 2,
else goto 4.
 ↔
- 4: If possible, expand **current partition** to the left, skim the **new partition** and regard the **new lexemes**,
else remember position as **left** boundary and goto 6.
 ← ?
- 5: If **current lexemes** \subseteq **new lexemes** then
 remember **new lexemes** as **current lexemes**,
 remember **new partition** as **current partition**
 and goto 4,
else remember position as **left** boundary and goto 6.
 ←
- 6: If possible, expand **current partition** to the right, skim the **new partition** and regard the **new lexemes**,
else remember position as **right** boundary and goto 8.
 → ?
- 7: If **current lexemes** \subseteq **new lexemes** then
 remember **new lexemes** as **current lexemes**,
 remember **new partition** as **current partition**
 and goto 6,
else remember position as **right** boundary and goto 8.
 →
- 8: No further expansion is possible from this starting partition. Return **left** and **right** boundaries of **current partition** along with **starting partition** positions.
- 9: If possible, shift the starting pair one noun to the left, accept this as **starting partition** and **current partition**, skim **current partition**, remember the lexemes of best interpretation as **current lexemes** and goto 2,
else goto 10.
- 10: No more untried starting partitions.
 Terminate.

Figure 15: The informal determinative algorithm for finding all scopes of unambiguity in natural language text. Note that expansion is only physical possible as far as the physical boundaries of S goes. Therefore expansion will be deemed impossible in lines 2, 4 and 6 if the respective physical boundaries of S have been encountered.

unambiguity all expand minimally. In such a sequence the only semantic coherence available will be the starting partitions themselves, as introduced a bit artificially by the assumption that two neighbouring nouns in a sequence probably goes together in the same scope of unambiguity. The scopes of unambiguity in such a sequence is shown in a). Such sequences must involve a series of polysemous nouns arranged in such a way that all engage a different meaning with each of their respective neighbours and thus they must be considered extremely rare under normal circumstances. At the other extreme, it may be the case that all the nouns in the sequence go together in the same scope of unambiguity as in b). No matter what pair of nouns in the sequence is chosen as a start partition it will expand to the same maximal scope of unambiguity bounded only by the physical bounds of the sequence itself. While sequences like the one in b), are probably more frequent in actual data than ones like a), any sequence over a certain length should probably be expected to involve more than one scope of unambiguity. Thus, the pattern of scopes in c) is likely to be representative of the majority of sequences from actual data, where several scopes of unambiguity are present while clearly distinguishable from each other. Here it should be an obvious idea to consider the overlap of scopes as a restriction on where the sequence might be meaningfully partitioned as indicated by the double line.

In chapter 3, I will apply this algorithm to the CIVIII corpus introduced in my thesis paper (Lassen, 2005) and point out the most

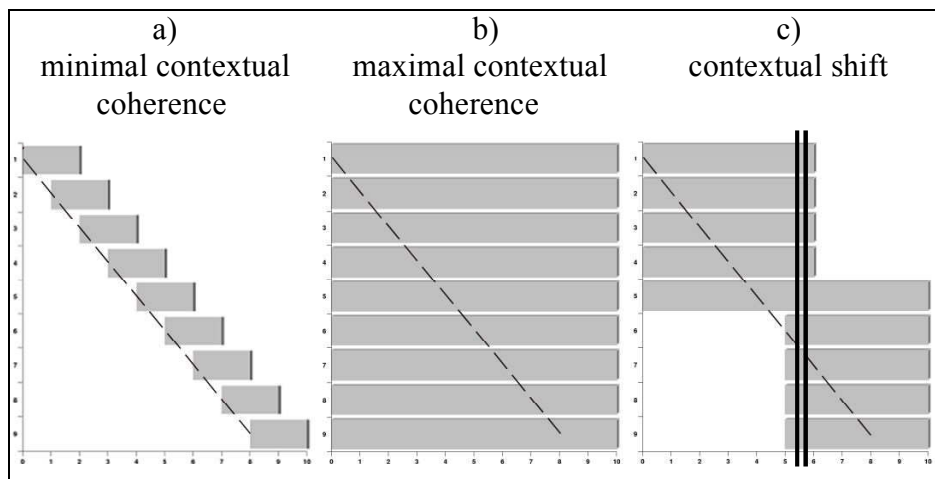


Figure 16: Three distinct patterns of scopes of unambiguity in a sequence of 10 nouns. The dotted diagonal indicate the position of first member of the starting partition of the respective iteration. In a) above is illustrated a case where no matter what starting partition is used, it can not be expanded unambiguously in any direction. The contrasting example in b) shows a case where all possible starting partition expands maximally without ambiguity. In c), I show a case in between these two extremes. Any variation of the pattern in c) should motivate consideration of whether to place a partitioning-line somewhere in the overlap of the scopes of unambiguity.

important issues to consider from the experimental results. I will also and suggest further experiments to be carried out.

3 Informal results and conclusions.

In the case of the small experimental CIVIII corpus we get 226 scopes. In order to best compare all those scopes, each will be represented by as a horizontal column stretching from its leftmost expansion to its rightmost expansion, keeping track of its respective starting-point. Doing this for all 226 scopes of the experimental corpus the combined columns occupy the grey areas of Figure 17. Regarding the way the scopes distribute over the span of S several observations seems worth noting.

Partitions show as squares of length relative to the number of instances in the partition, these squares of course place themselves along the dotted diagonal representing the starting point for the respective iteration. Scopes does, it seems, distribute over the experimental corpus in a way that suggest significant semantic segmentation of the text. Even though the skimming algorithm clearly need a lot of polishing, the semantic segmentation of the experimental text is clearly visible as consecutive grey squares along the diagonal of figure 17.

Furthermore, while it is perhaps not as apparent, the scopes of unambiguity does actually seems to align themselves to a certain degree with the paragraphs of the original text. There are several cases of grey “squares” echoing the black boundaries with a slight “delay” that can be explained as a consequence of the relative inaccuracy of the skimmer that I have discussed elsewhere. The algorithm simply takes a bit long to realize that a shift has taken place because. There are, however also several cases of borders of original paragraphs coinciding with boundaries of unambiguity scopes along either or both dimensions, more than mere coincidence can explain. This does suggest some very real and robust contextual turning-points throughout S that coincide with some of the paragraph boundaries placed by the author of the original text.

This being said, obviously there is far to much to examine as to why the graphical representation of figure 17 behaves as it does, than can be reviewed in a few short paragraphs. Far to much, as well, for the scope of this small paper. In order for this assignment to remain manageable, I will have to satisfy myself with the thorough analysis that I offered so far and postpone the real and exiting experiments for another occasion. I will therefore end this paper by summing up and point out some of the directions experiments could take.

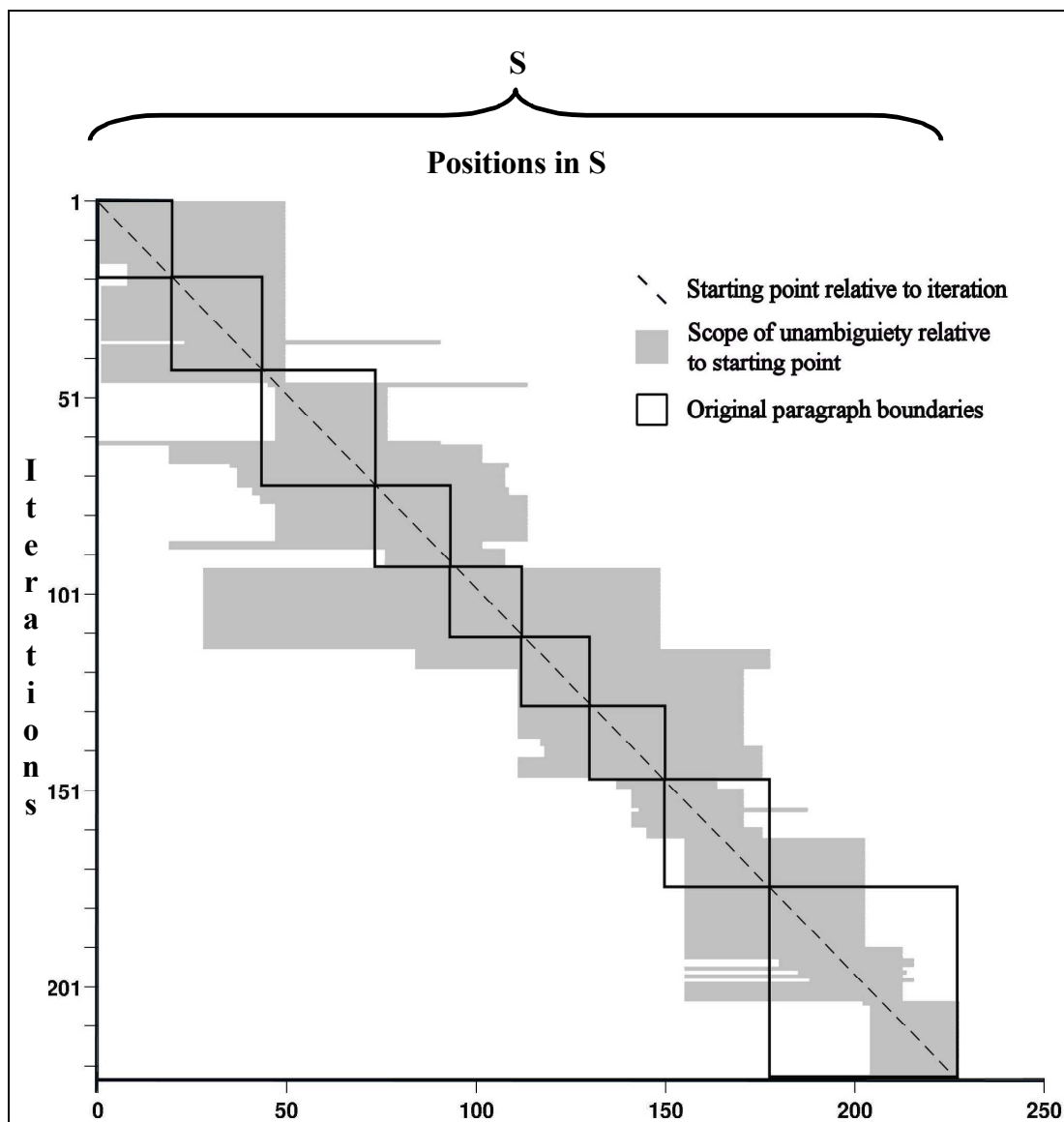


Figure 17: 226 overlapping scopes of unambiguity has been found in the CIVIII corpus. This representation gives a good overview of the entire sequence and variations in extension are easy to recognize. The dotted diagonal mark the beginning of the starting partition of the respective iteration while the square borders indicate the boundaries of paragraphs in the original text.

3.1 Conclusion

I set out to see if the experimental skimming algorithm could be applied to the automatic partitioning of running natural language text into meaningful and meaning-preserving portions. It seems, that the problem can be reduced to the proper distribution of squares along the diagonal in a figure like the one in **Figure 17**.

I have introduced the notion of “scope of unambiguity” and shown how it can be seen to play central role in the realization of contextual foci; one, that relates closely to the Gricean notion of cooperation between sender and receiver in language exchange.

I have introduced the distinction between “expansion and restriction” as approaches to finding partitions in noun sequences, and discussed their respective strengths and weaknesses thoroughly.

Having designed a working algorithm, applying skimming, that finds all scopes of unambiguity in a sequence of nouns, I have shown how scopes may be represented both graphically and formally in a robust and consistent manner.

Finally, regarding the experimental results in Figure 17, there is clearly a significant structure in the graphical representation of the distribution of the scopes of unambiguity. It is obviously possible - and even quite likely - that the graphical structure of Figure 17 reflects the semantic structure of the respective text. However, I have only managed to indicate that such a relation may be present, I have not proven its existence to any extent. An experimental algorithm was applied to an experimental corpus and the result does indicate soundness of theory while offering no proofs.

It is clear that more and larger experiments are necessary in order to be able to conclude anything further in this matter. In particular we still require a robust and consistent method of evaluation as well as measure of success, with regard to the quality of interpretations made by the skimmer and also the quality of actual partitions in a text. Such methods and measures quite likely involve the judgement of impartial human “test subjects”.

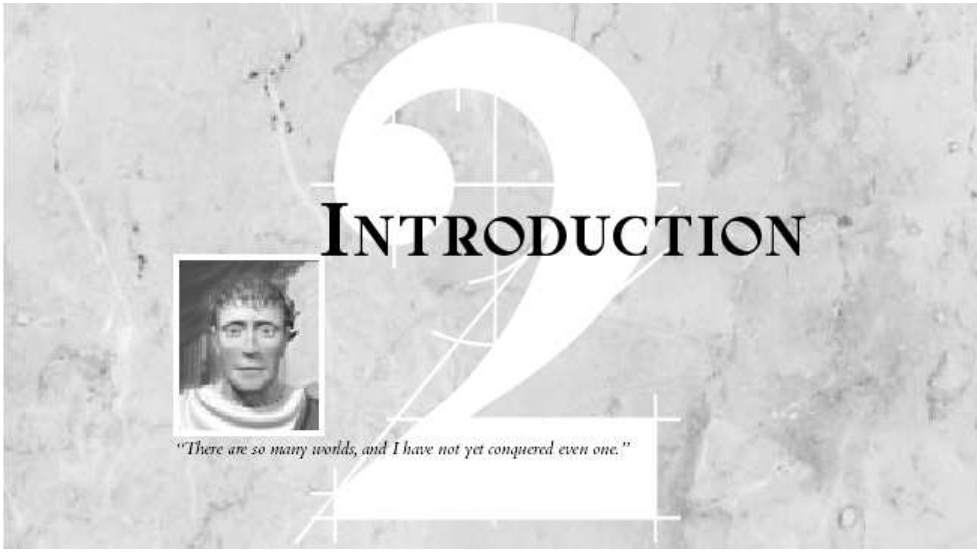
The analysis of this paper forms a firm ground and clear incentive for establishing an acceptable measure of success and subsequently a close study of the behaviour of the interpretations involved in this corpus. Also, extensive experiments with different corpora of varying size, of both real language and artificial data are necessary.

Appendix

A-1 References

Lassen(2005): “Skimming for Context” - Masters Thesis; Diku, University of Copenhagen

A-2 The experimental corpus, CivIII

	 The image shows the title screen for Civilization III. It features a large, stylized number '2' in the background. The word 'INTRODUCTION' is written in a serif font across the middle of the '2'. In the lower-left corner, there is a small portrait of a man with a beard and a white tunic. Below the portrait, the quote "There are so many worlds, and I have not yet conquered even one." is written in a small, italicized font. The background is a textured, light gray color.
0	<h3>Five Impulses of Civilization</h3> <p>There is no single driving force behind the urge toward civilization, no one goal toward which every culture strives. There is, instead, a web of forces and objectives that impel and beckon, shaping cultures as they grow. In the <i>Civilization III</i> game, five basic impulses are of the greatest importance to the health and flexibility of your fledgling society.</p>
1	<h3>Exploration</h3> <p>An early focus in the game is exploration. You begin the game knowing almost nothing about your surroundings. Most of the map is dark. Your units move into this darkness of unexplored territory and discover new terrain; mountains, rivers, grasslands, and forests are just some of the features they might find. The areas they explore might be occupied by minor tribes or another culture's units. In either case, a chance meeting might provoke a variety of encounters.</p>
2	<h3>Economics</h3> <p>As your civilization expands, you'll need to manage the growing complexity of its production and resource requirements. Adjusting the tax rates and choosing the most productive terrain for your purposes, you can control the speeds at which your population grows larger and your cities produce goods. By setting taxes higher and science lower, you can tilt your economy into a cash cow. You can also adjust the happiness of your population. Perhaps you'll assign more of your population to entertainment, or you might clamp down on unrest with a larger military presence. You can establish trade with other powers to bring in luxuries and strategic resources to satisfy the demands of your empire.</p>

3	<h3>Knowledge</h3> <p>On the flip side of your economics management is your commitment to scholarship. By setting taxes lower and science higher, you can increase the frequency with which your population discovers new technologies. With each new advance, further paths of learning open up and new units and city improvements become available for manufacture. Some technological discoveries let your cities build unique Wonders of the World.</p>
4	<h3>Conquest</h3> <p>Perhaps your taste runs to military persuasion. The <i>Civilization III</i> game allows you to pursue a range of postures, from pure defence through imperialistic aggression to cooperative alliance. One way to win the game is to be the last civilization standing when the dust clears. Of course, first you must overcome both fierce barbarian attacks and swift sorties by your opponents.</p>
5	<h3>Culture</h3> <p>When a civilization becomes stable and prosperous enough, it can afford to explore the Arts. Though cultural achievements often have little practical value, they are frequently the measure by which history—and other cultures—judge a people. A strong culture also helps to build a cohesive society that can resist assimilation by an occupying force. The effort you spend on building an enduring cultural identity might seem like a luxury, but without it, you forfeit any chance at a greatness other civilizations will respect.</p>
6	<h3>The Big Picture</h3> <p>A winning strategy is one that combines all of these aspects into a flexible whole. Your first mission is to survive; your second is to thrive. It is not true that the largest civilization is necessarily the winner, nor that the wealthiest always has the upper hand. In fact, a balance of knowledge, cash, military might, cultural achievement, and diplomatic ties allows you to respond to any crisis that occurs, whether it is a barbarian invasion, an aggressive rival, or an upsurge of internal unrest.</p>
7	<h3>Winning</h3> <p>There are now more ways of winning the game. You can still win the Space Race with fast research and a factory base devoted to producing spacecraft components. You can still conquer the world by focusing on a strong military strategy. If you dominate the great majority of the globe, your rival may well give in to your awesome might. In addition, there's a purely Diplomatic means of success; if you're universally renowned as a trustworthy peacemaker, you can become head of the United Nations. Then there's the challenge of overwhelming the world with your Cultural achievements—not an easy task. Finally, of course, is perhaps the most satisfying victory of all—beating your own highest Isographic Civilization Score or those of your friends. See Chapter 14: Winning the Game for an in-depth analysis of the scoring system.</p>
8	<h3>The Documentation</h3> <p>The folks who make computer games know that most players never read the manual. Until a problem rears its head, the average person just bulls through by trial and error; it's part of the fun. When a problem does come up, this type of player wants to spend as little time in the book as possible, then get back to the game. For those of you who are looking for a quick fix, Chapter 15: Reference: Screen by Screen is the place to go. For the rest of you, we've tried to organize the chapters in the order that you'll need them if you've never played a <i>Civilization</i> game before. If you're new to the game, the sidebars on concepts should help you understand the fundamentals of the game. The Readme file on the CD-ROM has the rundown on the very latest changes, things that didn't make it into this manual. (Due to printing and binding time, the manual has to be completed before final tweaks are made.) Last but not least, the <i>Civilization III</i> game continues the tradition of including a vast compendium of onscreen help. Click on the Civlopedia icon (the book near your advisors) or on any hyperlinked text in the game to open the Civlopedia. This handy reference includes entries describing all the units, improvements, governments, terrain, general game concepts, and more—everything you could want to know about the <i>Civilization</i> world. The entries are hyperlinked so you can jump from one to another with ease.</p>

