Université de Montréal

Linguistic Processes for Content Condensation in Abstracting Scientific Texts

par

Choy-Kim CHUAH

Département de linguistique et de traduction

Faculté des arts et des sciences

Thèse presentée à la faculté des études supérieures

en vue de l'obtention du grade de

Philosophie Doctor (Ph.D.)

en linguistique

avril 2001

Université de Montréal

Faculté des études supérieures

Cette thèse intitulée

Linguistic Processes for Content Condensation in Abstracting Scientific Texts

presentée par

Choy-Kim CHUAH

a été evaluée par un jury composé des personnes suivantes

xxx

Richard PATRY

(président-rapporteur)

xxx

Richard KITTREDGE

(directeur de recherche)

xxx

Kathleen CONNORS

(membre du jury)

xxx

Elisabeth LIDDY (Représentant Jean-Yves MORIN)

Université de Syracuse (Université de Montréal)

(examinateur externe)

Thèse acceptée le: 27 août 2001

Summary

While *content selection* has been intensively explored in the sentence extraction approach to automatic summarization, there is generally little work on the other process of *content condensation.*

To understand this process of condensation, we propose a partial typology based on whether a linguistic unit is replaced, deleted, compressed into fewer essential units, or combined with another unit. Four important categories of condensation processes: *generalization*, *deletion*, *compression*, and *aggregation*, including their inverse processes, e.g. *insertion*, and *expansion*, which were occasionally observed, are proposed. To guide the usage of the same term for similar operations, we borrow definitions from linguistics. The type and function of the linguistic units involved are also discussed. We carried out an empirical analysis of 57 author-written abstracts of on-line journal articles in entomology, tracing each abstract sentence back to the plausible source sentences in the corresponding full text. Unlike other studies which focus on the resultant abstract, our study focuses on the processes leading to the production of abstract sentences from corresponding full-text sentences. We do not, however, propose an algorithm for abstracting, or account for all the conditions under which individual condensation operations may apply.

While a range of substitutes were used in abstracting, about half of the stems of lexical units in our abstracts share the same stem as their source words, or are their derived forms. Only a small proportion of substitutes were synonyms, and the rest were (quasi-)synonyms, or imprecise equivalents. Authors tend to use less technical forms in abstracts possibly in anticipation of non-specialist abstract readers. Numerical expressions are rendered less precise although no less accurate: absolute numbers and decimals are rounded off, and percentages replaced by ratios or fractions. These observations are consistent with the "new" context of an abstract where only the gist of a document's content need be re-conveyed.

Among the linguistic units commonly deleted are metadiscourse phrases, and segments of text (e.g. parenthetical texts, and apposed texts), which provide details and precision in the full

text, but are out of place in an abstract. Redundancies inserted for various reasons, or units deemed to be implicit to the comprehension of targeted readers are also often removed. While deletion is an important sub-process of condensation, we observed some instances of adding experimental and other details to compact more information into abstract. The expansion or "unpacking" of compact linguistic units was also observed. The secondary role of inverse processes observed calls for a review of the meaning of condensation from "not giving as much detail or using fewer words" to include the adding of information in order to make a unit of text informatively compact.

Among the linguistic units compressed are verbal complexes containing a support verb, or a catenative. Like semantically empty support verbs (e.g. *X caused decreases in Y = X reduced Y*), some catenatives too may be deleted without significant changes in meaning to the verbal complex (e.g. *X was allowed to hatch* ≅ *X hatched*). Redundancy in meaning between an adjective and a noun in a noun phrase, e.g. *functional role*, may be removed, and the phrase compressed to just the stem of the adjective, i.e. *function.* While not frequently occurring in the corpus studied, the compression of such units may be described by rules, and hence, might be operationalized for automatic abstracting.

Aggregation, the combining of units of text within or between sentences, is an important sub-process of condensation. Two-thirds of sentences in abstracts studied were written using multiple sentences, and more sentences were combined without than with the use of an explicit sign, such as a connective, a colon or a semi-colon.

If research in summarization is to progress beyond sentence selection, then we must work towards: (a) a clear distinction between operations that are condensation processes, and those that are not; (b) bringing operationally similar processes together under the same designation, and (c) a greater understanding of sub-processes constituting condensation. To this end, our provisional typology for condensation, the range of type of linguistic units involved and their functions sets the first step to advance research into content condensation. We have only just

begun to identify the condensation sub-processes in operation during abstracting. The factors that are critical on the interplay of these processes still need to be investigated.

Keywords: abstracting, condensation, substitution, deletion, metadiscourse

Résumé

Les approches d'extraction du texte pour faire un résumé par ordinateur ont surtout porté sur la sélection du contenu. La condensation du contenu n'a presque jamais été traité.

Afin d'explorer le processus de condensation, nous proposons une typologie partielle basée sur des manipulations. Une unité lexicale peut-être combinée avec une autre unité lexicale, remplacée, effacée ou compactée. Pour la plupart de ces processus (i.e. agrégation, généralisation, effacement et compression) et les processus inverses, comme l'insertion et l'expansion, nous empruntons les définitions de linguistique qui a pour but de guider l'usage les désignations proposées. Le type et la fonction des unités lexicales impliquées sont aussi discutés. N'ayant pas accès aux processus de résumé, nous les avons déduits en comparant les phrases constituant un document et le résumé correspondant. Contrairement à d'autres études qui focussent sur le produit, c'est-à-dire le résumé, nous nous concentrons sur les processus de production sans toutefois produire de résumé. Nous ne proposons pas un algorithme pour faire un résumé, et nous ne essayons pas d'éxpliquer tous les conditions en opération. Cette étude porte sur 57 articles en entomologies dont les résumés ont été préparés par l'auteur. Ces articles se trouvent en ligne.

L'environ la moitié des unités lexicales utilisées dans un résumé ont les mêmes lexèmes que les unités lexicales du document source ou en sont dérivées. L'utilisation des mêmes lexèmes garantit qu'on parle du même concept, et qu'il n'y a pas de changement de sujet. Peu d'unités lexicales du résumé est des synonymes des unités lexicales du document. Le reste des unités lexicales du résumé sont des les quasi-synonymes ou des équivalents imprécis. Les auteurs ont tendance à utiliser dans un résumé les formes non-techniques probablement en visant des lecteurs non-spécialistes. Les expressions numériques sont mois précises mais exactes. Les nombres absolus sont arrondis et les pourcentages sont remplacés par des proportions et des ratios. Ces observations sont en accord avec le nouvel context d'un résumé où on garde que l'essentiel du document.

Parmi les unités lexicales linguistiques qui sont souvent effacées, on retrouve les metadiscours et les segments du texte qui donnent les détails et les précisions comme, par exemple, les textes entre parenthèses et les textes apposés. On élimine aussi les redondances ou les unités lexicales implicites. Même si l'effacement est lié à la condensation, on ajoute parfois des détails expérimentaux pour mieux compacter l'information ou on réalise l'expansion d'unités lexicales condensées. Le rôle secondaire de ces processus inverses observés impose qu'on revoit la signification de la condensation qui au sens général signifie 'donner le moins de détail possible ou utiliser moins de mots' pour inclure un ajout d'information afin de rendre un résumé uniformément plus compactes.

Parmi les unités lexicales linguistiques compactés, on a également relevé des syntagmes verbaux contenant un verbe du support ou un « catenative ». Le verbe du support dans un syntagme verbale est effaçable sans grand changement sémantique au sens du syntagme (par exemple, *X caused decreases in Y = X reduced Y*). Comme les verbes supports sont sémantiquement vides, certains « catenatives » sont aussi effaçables (par exemple, *X was allowed to hatch* ≅ *X hatched*). Une redondance sémantique entre un adjectif et un nom dans un syntagme nominal, par exemple, *functional role*, est aussi effaçable, pour obtenir le lexème de l'adjectif, c'est-à-dire *function*. Malgré la faible occurrence de ces unités lexicales dans notre corpus, nous pouvons les décrits en utilisant des règles opérationnelles. Néanmoins, les conditions d'opération de ces règles ont besoin des études plus cherchées.

La combinaison des unités du texte dans une phrase ou entre des phrases (agrégation) est un processus important de la condensation. Les 2/3 des phrases du résumé ont été écrites en utilisant des phrases multiples. La plupart du temps, pour faire une phrase du résumé, nous n'utilisons pas un signe explicite (un connecteur, un deux point ou point virgule) pour combiner les éléments des phrases du texte source.

Pour avancer la recherche sur la production des résumés, il est nécessaire de : (a) distinguer clairement les processus les vrais opérations, c'est-à-dire les vraies opérations de condensation ; (b) donner la même désignation aux mêmes processus; et (c) obtenir une meilleure compréhension des sous-processus qui constituent la condensation.

Note typologie provisoire de la condensation, ainsi que la gamme de type et la fonction des unités lexicales impliquées (par exemple, pour les processus du remplacement et de l'effacement) ont permis de faire avancer la recherche en condensation du contenu. Nous avons identifié les sous-processus de la condensation, mais plusieurs facteurs qui influencent l'interaction méritent des études plus poussées.

List of Tables

List of Figures

List of Abbreviations

ϕ	omission
⌈text⌉	text to be treated as a single unit
¬WN	not in WN
A	adjective
ab-	abstract
ADJ	adjective
Adv	adverb
$Adv_0(A)$	adverb derived from adjective
Conj	conjunction
~~deleted text~~	deleted text has a ~~strikethrough~~
Det	determiner
ECD	Explicatory-Combinatorial Dictionary
ft-	full text/document
Gener(ic)	ECD lexical function for generic word that can occur syntagmatically
LF	lexical function
LU	lexical unit
LU_{head}	syntactic function of lexical unit is a head
LU_{modf}	syntactic function of lexical unit is a modifier
modf	modifier
N	noun
NP	nominal phrase
PARTCL	participial clause
Paren-txt	parenthetical text
PP	prepositional phrase
PREP	preposition
Qtf	quantifier
RCL	relative clause
S	sentence
S' (or S" or S"')	modified sentence
$S_0(V)$	substantive form derived from verb
SCL	subordinate clause
SR	special lexical resource
Syn(onym(y))	ECD lexical function for synonymy relation
text added on	text added on is underscored with a thick line
V	verb
VP	verbal phrase
WN	WordNet
X, Y, or Z	unspecified text segments

Acknowledgement – Remerciements – Penghargaan

The completion of any project depends on the combined effort of many people each playing a different role. Special thanks go to the publishing houses, abstracting services (Chemical Abstracts Services and BIOSIS) and busy individuals who took the time to answer my numerous questions and even send me material. I would like to thank Springer Publications for permission to reprint bits of texts from documents downloaded for the study. While it is not possible for me to acknowledge everybody, I would like you to know that all of you have a special place in my heart.

To begin, I would like to thank my supervisor, Dr. Richard Kittredge, for his guidance throughout the project. Thanks for watching out for me.

J'aimerais remercier les professeurs du département de linguistique et traduction, en particulier M. Alain Polguère et encore plus particulièrement M. Igor Melcuk. J'aimerais que vous sachiez M. Mel'cuk, que vous étiez mon cauchemar autant j'étais le vôtre! Grâce à votre enseignement, je vois la morphologie différemment. Merci mille fois!

Au personnel du secrétariat du département de linguistique et aux étudiants et collègues du département, je vous remercie pour votre amitié et votre soutien. À Lyne et Eliana, merci.

J'aimerais aussi remercier les professeurs et les étudiants du département d'informatique et de recherche opérationelle, notamment M. Guy Lapalme, M. Jian-Yun Nie et Horacio. *Muchas gracias por tu ayuda, has sido muy amable conmigo.*

À l'Université de Montréal, je vous remercie pour l'aide financière.

À ma famille à Montréal, merci.

Kepada Universiti Sains Malaysia diucapkan ribuan terima kasih atas biasiswa dan perlanjutan demi perlanjutan cuti enam bulanan yang diberikan.

Kepada Zarin, Kilroy mengucapkan terima kasih.

To Tina, Kenny, Sodhy, Chooi Heong, Sim and Boni, thanks for running my numerous errands. To Maria *efharisto* for your moral support in the early part of the project, and your help at the end. And, to the convent sisters, I know that I am in your daily prayers. Thank you and Him.

Last but not least, a rare expression of my love to my family for their unquestioning support in my decision to go back to school, and thank you for taking good care of yourselves and each other that I may be able to pursue my studies without worry.

Mum, dad, this doctorate is dedicated to both of you. Look at what both of you have achieved? And dad, you can't even read? Mum, thank you for instilling in us the value of education, and for picking out the red grades from the blue ones in our report cards. Dad, thanks for making sure each evening that we have sharpened pencils for school the next day. To my children, I urge you not to take the path that I took, but to follow your heart as I did mine.

This doctorate goes as much to each and everyone of you as it does to me. I would not have been able to complete it without your help. You do not know how important a role all of you have played in my life. To all of you my deepest respect and love.

To Life itself, I cannot begin to express my gratefulness for the numerous second chances. I asked for only one extra day to be able to take a walk, and to prove them wrong, but you gave me so many that I have stopped counting. The fortitude gained from the gift of that experience has sustained me through these years. And, I dread to think what you have in store for me.

K

WHEN NOTHING SEEMS TO HELP, I GO AND LOOK AT A STONECUTTER HAMMERING AWAY AT HIS ROCK PERHAPS A HUNDRED TIMES WITHOUT AS MUCH AS A CRACK SHOWING IN IT. YET AT THE HUNDRED AND FIRST BLOW IT WILL SPLIT IN TWO, AND I KNOW IT WAS NOT THAT BLOW THAT DID IT – BUT ALL THAT HAD GONE BEFORE.

Jacob Riis

Contents

Chapter 1

Introduction and Motivation

In Alexandrian times, scholars would write a short description of a document on a tag and attach it to the scroll to help retrieve the relevant parchment without having to unroll it. Thus began the tradition of attaching a summary to a document to describe its content.

Until recently, the type of documents for which summaries were written was primarily academic. Today, with the advent of the Web as a means of communication and information transfer, not only do we have journals on-line, but also writings on a wide range of political, social or economic subjects. While summaries are required to help a reader keep abreast with new developments in non-academic writings which have a high turnover, abstracts are required for academic articles that are produced at a rate of a few thousand per day.

In the year 2000, *Chemical Abstracts Services* (CAS), the world's largest secondary information service in chemistry and chemical engineering, abstracted and indexed 725,195 published documents, including patents, books and papers (journal articles as well as dissertations, technical reports, and conference proceedings). This rate which has increased exponentially over the last century (see Fig. A1-1 in Appendix I), is now close to two thousand per day. For *BioSciences Information Services* (*BIOSIS*), a major publisher of biological abstracts, the rate is about a thousand per day[1]. While both services may depend heavily on authors of papers (nearly 90% of records for *Biological Abstracts* are prepared by an author himself), abstracts still need to be written for the other documents.

Chemical Abstracts Services (CAS) covers journals from nearly 200 countries and in about 50 languages. While about 75%[2] of original documents (including patents) abstracted and indexed are published in English (pers. comm. from CAS Help Desk; 13 March 2001), translation is additionally required for documents that are not in English. In the year 2000, 45.2% of patents abstracted by CAS were in Japanese. Although no recent figure on the cost of writing

[1] "*Biological Abstracts* includes approximately 350,000 accounts of original research yearly from nearly 6,000 primary journal and monograph titles" (from http://library.dialog.com/bluesheets/html/bl0005.html. See also, http://www.biosis.org/pdfs/BAfact.pdf). BIOSIS is reported to have 213 employees.

[2] Or, 82.9% if papers alone are considered.

an abstract is available[3], an abstract was reported in 1968[4] to cost CAS about $23-25 (inclusive of research budget) to produce.

This need for summaries, and quickly, has fueled interest in automatic summarization in the last decade. The 1993-2000 period alone saw five conferences[5], and even one initiative on its evaluation[6]. Our present research is drawn into this "frenzy" in automatic abstracting. While most studies focus on the product, we will concentrate on the processes leading to its production where research is very much needed.

In the most studied of approaches in automatic summarization, sentence extraction, summarization proper may be divided into two processes of CONTENT SELECTION and CONTENT CONDENSATION. While selection has been intensively explored, there has been in general little work on the second process of condensation, specifically studies that identify the processes involved in the condensation of selected sentences, or look into their interplay.

The earliest work which we are aware of was carried out about three decades ago[7]. However, in the past year interest on condensation has resurfaced. Some operations/ transformations to edit extracted sentences to produce concise texts were identified. The designations proposed for these processes however lack general consensus. Linguistically similar operations may be designated differently by independent authors, and some operationally similar processes may be designated differently by the same author. Discrepancies in designations are not conducive to advancement in summarization.

---

[3] To our request for a recent figure, CAS is unable to respond as the "financial information is the property of the American Chemical Society".

[4] In the Encyclopedia of Library and Information Science (Kent & Lancour, 1968).

[5] (a) The ANLP/NAACL 2000 Workshop on Automatic Summarization in Seattle, Washington;
(b) The *Rencontre Internationale sur l'extraction le Filtrage et le Résumé Automatique* (RIFRA 98) in Tunisia;
(c) The AAAI 1998 Spring Symposium on Intelligent Text Summarization in California; and
(d) The ACL'97/EACL'97 Workshop on Intelligent Scalable Text Summarization in Spain.
(e) The Dagstuhl Seminar on Summarising Text for Intelligent Communication held in Germany in December 1993, was said to be the "first wholly devoted to automatic summarizing" (Sparck Jones & Endres-Niggemeyer, 1995:625).

[6] Namely, the TIPSTER SUMMAC Text Summarization Evaluation in 1998 organized by the U.S. Government. The initiative involved sixteen participants from four countries (Mani *et al.*, 1998:1).

[7] See Mathis, Rush & Young (1973).

Consider operations/transformations (a), (b) and (c) which are all essentially deletion processes.

(a)	~~When it arrives,~~ X	→ X'	(author-1)
(b)	~~In this paper we~~ have presented X	→ Presents X';	(author-2)
(c)	~~In this paper, we~~ report X	→ Reports X';	(author-2)

While author-1 designated operation (a) as SENTENCE REDUCTION, author-2 labeled operation (b) as SYNTACTIC VERB TRANSFORMATION (where X is a segment of text, and X' is its modified form). At the same time, transformations (b) and (c), which are essentially the same, were designated differently as SYNTACTIC VERB TRANSFORMATION and CONCEPTUAL DELETION by the same author. We remind that the focus of the authors in question was on the product, and not on the processes leading to its production.

Besides discrepancies in designations, some operations identified are in fact manipulations consequent from processes implicated in the writing of abstracts, e.g. moving a subject from the end to the front, and acronym expansion. While such operations were not unequivocally stated to be non-condensation processes, the operations were given and discussed together with operations that were. If research in summarization is to be encouraged beyond sentence selection, then: (a) a clear distinction between operations that are condensation processes, and those that are not, is prerequisite; (b) operationally similar processes need to be brought together under the same designation, and (c) a greater understanding of sub-processes constituting condensation needs to be obtained. To this end, a typology for condensation, their designations and definitions need to be established.

The need for automatic summarization has long preceded the opportunity to do that which could not be done before. Today we have at our disposal a variety of on-line text corpora, and linguistic tools. On-line dictionaries and thesauri serve as rich sources of linguistic information ready to be exploited, while on-line journals and web-sites serve as ready text corpora for the investigation and refinement of linguistic tools. The present research responds to the need to understand condensation by identifying the linguistic mechanisms implicated in the

writing of abstracts for scientific journal articles on biology where growth has been exponential[8]. This lack of knowledge on condensation probably explains in part why research on automatic summarization has been held back at selection.

1.1 Aims of Study

In contrast to current research in automatic summarization which has mainly focussed on content selection to get to the final product of a summary, our research seeks to understand the process of content condensation by identifying the linguistic mechanisms implicated in abstracting, and the range of types of the linguistic units involved and their functions. By (a) setting out designations proposed by various researchers in summarization, (b) reviewing some linguistic mechanisms in concise re- expression of content; and (c) identifying processes in writing abstracts by an author, we look into a provisional typology for condensation, which is a necessary first step towards the consorted use of terms, and a definition of condensation and its sub-processes in summarization in general. It is not in the aim of the present study to produce any abstract, nor to look into the interplay of the processes.

Briefly, our research which was carried out on a scientific corpus on entomology-related articles, seeks:

1. To determine the linguistic mechanisms implicated in condensation;
2. To propose, on the basis of the above findings, a definition for a linguistically-based typology for condensation; and
3. To identify the range of types of linguistic units involved and their functions;
4. To explore the utility of WordNet to summarization in general, and abstracting in particular.

From data obtained from the above study, we show why some commonly-used features are unreliable as basis of sentence selection at least for the scientific corpus investigated.

[8] "In general, the number of scientific, scholarly periodicals has doubled every ten to fifteen years during the twentieth century. ... The increase during the last decade has been especially explosive for the life sciences, making it one of the fastest-growing disciplines" (Davis & Schmidt, 1998).

1.2 Organization of Document

The rest of this document is divided as follows into nine chapters.

Chapter 2 which provides an introduction to automatic summarization, begins by presenting definitions proposed for the two terms SUMMARY and ABSTRACT, before going on to present Sparck Jones' three-stage process of summarization, and three current techniques in sentence extraction for summarization. Next, we categorize the designations identified or proposed for processes in content condensation by fellow researchers in summarization and text generation. The categorization was based on whether a linguistic unit is: (a) substituted, (b) deleted or added on to, or (c) combined with or separated from another unit. The chapter ends with a look at some fields of research potentially contributing to abstracting.

Chapter 3 describes the methodology for the study of linguistic mechanisms in content condensation, the matching process and some difficulties encountered. In Chapter 4, we provide some statistics on sentences identified to have been selected for abstracting. The distribution obtained sheds light on the sections where sentences for abstracting are likely to be located, and in what proportions. While indicative of the potential use of text structure as a possible feature in abstracting structured scientific documents, the distribution of selected sentences also sheds light on the significance of aggregation in condensation. From sentences identified to have been used in abstracting and the statistics obtained, we are able to confirm the reliability of some cues commonly used in sentence selection in current techniques in summarization.

In Chapter 5, we look at some linguistic mechanisms the English language has for concise reformulation of content, before going on to determine the processes actually used by authors in writing abstracts for scientific articles. Without access to how authors abstract, the processes were deduced indirectly via a comparative study of information in abstract and their corresponding sources in full text sentences. Based on the categorization of designations for condensation processes discussed in Chapter 2, condensation mechanisms from linguistics and from our comparative study, we end the chapter with a typology of condensation sub-processes in abstracting. A fourth category was introduced for linguistic units there are expanded, or reduced to its essential units. For each of the four groups of processes:

(a) generalization,

(b) deletion,
(c) compression, and
(d) aggregation,

we propose definitions adapted from those available in linguistics or from those proposed by fellow researchers. The inverse operations of these processes are also discussed.

Chapters 6, 7, 8, and 9 report the linguistic mechanisms identified for condensation in abstracting, namely substitution, deletion, compression, and aggregation, and the linguistic units involved.

Chapter 10 sums up the findings and orientates the direction of future research.

Chapter 2

Automatic Summarization: An Introduction

The ultimate aim of summarization is to produce a SUMMARY or an ABSTRACT. Chapter 2 begins by giving the definitions for the two terms, as proposed by the American National Standards Committee (ANSC) Z39 (1979), and the definition for summary from the domain of automatic summarization itself. In section 2.2, we present Sparck Jones' (1999) three-stage process of summarization.

The most explored of approaches in automatic summarization is that of SENTENCE EXTRACTION or SELECTION. Section 2.3 discusses three current techniques in sentence selection: (a) frequency/statistical technique; (b) lexical cohesion technique, and (c) Rhetorical Structure Theory technique. While extracted sentences may be further edited to condense them, research has generally stopped here.

Apart from the work of Mathis, Rush & Young (1973), and two recent studies, there has been little research into the process of content condensation. Using the definition of condensation in general language as basis, section 2.4 sets out the differing designations proposed by these researchers and others from summarization and text generation. The provisional categorization is an important first step to a typology of condensation sub-processes which is prerequisite to a consorted use of terms.

During summarization, content is re-expressed not just using linguistic units of different forms. The units used may share the same stem (e.g. *differences* → *differ*), be linguistically related (e.g. *dissect* → *cut*), or involve world knowledge (e.g. *August* → *summer*) or domain knowledge (e.g. *butterfly* → *Lepidoptera*). For this, we discuss in the last section some fields which are potentially contributing to summarization, namely lexicology and sublanguage.

2.1 A Definition: Summary vs. Abstract

Amongst the words meaning "result of text reduction", ABSTRACT and SUMMARY are the two most frequently used, often interchangeably. The ANSC Z39 (1979) explicitly defined the two terms to be different within the context of scientific and academic writing (for the complete definitions of abstract, abstracting and summary see Appendix II).

> ABSTRACT: an abbreviated accurate representation of a document, without added interpretation or criticism and without distinction as to who wrote the abstract.

> SUMMARY: a restatement within a document (usually at the end) of its salient findings and conclusions, and is intended to complete the orientation of a reader who has studied the preceding text.

The ANSC asked that the terms "not be used synonymously", nor should an abstract be called a summary. The reason given is because an abstract contains information from "vital portions of the document (for example, purpose, methods)", not found in a summary. While the ANSC may consider Method to be vital, our study showed this section of document to be the least important in terms of information extracted for abstracting (see Fig. 4-2). The question raised here is from which sections to extract information to write abstracts for structured technical texts? The sections as prescribed in the definition of an abstract by ANSC? Or, the sections where authors are most likely to place them? Or, wherever they might be found, so long as the information fulfills the criteria for selection specified? The choice depends on the abstractor and the abstracting situation.

Unlike the ANSC, Sparck Jones (1999) who had automatic summarization in mind, did not make any distinction between the two terms. Her provisional definition is grounded on the process(es) leading to its realization, rather than what it should contain. Summarizing is taken to include "extracting, abstracting, etc."

> SUMMARY: a reductive transformation of source text to summary text through content reduction by selection and/or generalisation on what is important in the source.

In the definitions of summary and abstract, the ANSC made precise reference to the "portions of the document" from which they are to be constituted. We make no such insistence, nor will attempt to further define the two terms. Henceforth, we will use the term ABSTRACT to refer to a special kind of summary associated with scientific and technical documents (excluding manuals[9]). Where we do not wish to draw any attention to the type of document, term SUMMARY will be used; abstract is a hyponym of summary.

2.2 The Automatic Summarization Process

By taking summarization to subsume abstracting, it is implied that the process of summarization described, and the factors affecting summarization and its evaluation, apply as well to abstracting.

In Sparck Jones's (1999) review of automatic summarization, she emphasized the importance of breaking the process of summarization down into distinct stages to allow "for checking the real logic underlying specific systems, making it easier to identify the assumptions on which they are based, and to compare one system with another." We diagram in Fig. 2-1 the three-stage process. Each stage is assumed to be further divisible.

(a) interpretation of source text content to arrive at a source text representation;
(b) transformation of source text representation into summary representation; and
(c) generation of summary from summary representation.

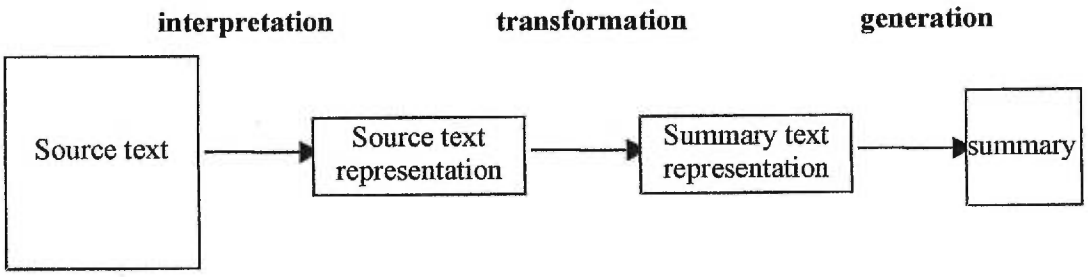


Fig. 2-1. Automatic Summarization Process

[9] For which abstracts are not normally required.

During the first stage, the source text is interpreted, first 'locally' at the level of individual sentences before being integrated 'globally' to give the source text representation (Sparck Jones & Endres-Niggemeyer, 1995:627). During the second stage of the summarization process, the source text representation is transformed into the summary text representation.

The third and final stage in summarization is the generation of a text summary from the summary representation. The amount of "smoothing" required for the summary representation to arrive at the summary is dependent on the technique used. While she did not specify which approach it was for, the model is most pertinent to that by sentence extraction[10]. Some techniques in sentence extraction, e.g. rhetorical structure relations technique (see section 2.3.3.), produce more readable representations than others, e.g. frequency/statistical technique (see section 2.3.1).

2.3 Techniques in Content Selection

There are three current techniques for extracting important sentences. The first technique which is statistics-based makes use of a mix of word occurrence, text cues and other features without taking into consideration the meaning of words. The second and third techniques are linguistics-based. While one exploits the relation that exists between words, and the other, the rhetorical relation that holds between segments of clauses. We describe each of them below.

2.3.1 Frequency/Statistical Technique

Text surface cues are the most obvious indication of text content. Hence, it is not surprising that word frequency, an indication of the topic discussed, was the earliest of techniques used (Luhn, 1968). However, because the significance of a word is not just its frequency in the document, the use of a simple word count as the basis for selecting important sentences to constitute a summary cannot be satisfactory as a summarization technique. Frequency count is the simplest of this category of techniques.

[10] The other two approaches to summarization: (a) template instantiation, and (b) generation, will not be discussed.

As refinements to the technique, some researchers (Edmundson, 1969; Rush *et al.*, 1971; Kupiec *et al.*, 1995) use a mix of cue words (e.g. *greatest*, *significant*), indicator phrases (e.g. *In this study*, *Our report indicates that*) and location in text (i.e. within document or within sentence) to enhance the formula in the extraction procedure. While cue words can be spurious in signaling important text material, the use of indicator phrases is related to writing style and text genre, and hence, bear on the technique. While Luhn (1968) exploited the proximity of significant words to each other, Brandow *et al.* (1995) used both document and corpus word frequency to identify the words that are unique to a document. Baxendale (1958:354) who noted the high probability of topic sentences being in initial and final positions (85% and 7% respectively), investigated the extraction of vocabulary from such sentences for indexing.

Statistical techniques do not take into consideration the semantic continuity of text. As a result the summary is disjointed, even if the sentences are themselves complete. However, for want of good summarization techniques, and the ease at which data can be manipulated, statistical techniques continue to be the most practiced[11]. The algorithm whose formula is able to pick out important sentences with the greatest of probability produces the best extracts.

Below, we describe Kupiec *et al.*'s (1995) document summarizer which represents a current standard of the state of the art.

A Trainable Document Summarizer - Kupiec, J., Pedersen, J. & Chen, F. (1995)

The work of Kupiec *et al.* was aimed at producing an extracted summary that is intermediate between full text and title, i.e. sufficiently informative to act as surrogate, yet short enough to be perused at a single glance, like a title. Summary sentences were extracted based on the basis of their probable significance. After experimenting with several features, the following five were chosen: (a) sentence length, where sentences shorter than a particular length, e.g. title, headings, were excluded; (b) fixed-phrase, e.g. *In conclusion*; (c) paragraph, i.e. location of sentence within paragraph; (d) thematic word, i.e. if open class word is frequently occurring, and (e) uppercase word, i.e. if word is a proper name. Of the features, the first three gave the best results. Poor results obtained with the thematic word and uppercase features were attributed to the fact

[11] "Work presented at the 1997 ACL Workshop on Intelligent Scalable Text Summarization primarily focused on the use of sentence extraction." (Radev & McKeown, 1998:473).

that such words were found throughout the text. However, because of their robustness, and the need to include more dispersed informative material, these two features were retained. We note that amongst the features, fixed-phrase, paragraph, and uppercase word are not entirely genre and style independent. This prevents free applicability of technique to any document type. An evaluation was carried out by comparing the sentences selected by the automatic summarizer against those picked by professional abstractors. The summarizer was reported to be capable of extracting up to 84% of the sentences chosen by professional abstractors. Kupiec *et al.* acquired their study corpus from Engineering Information Co., a non-profit company, which provide abstracts on technical articles to information services. Fig. 2-2 illustrates the process.

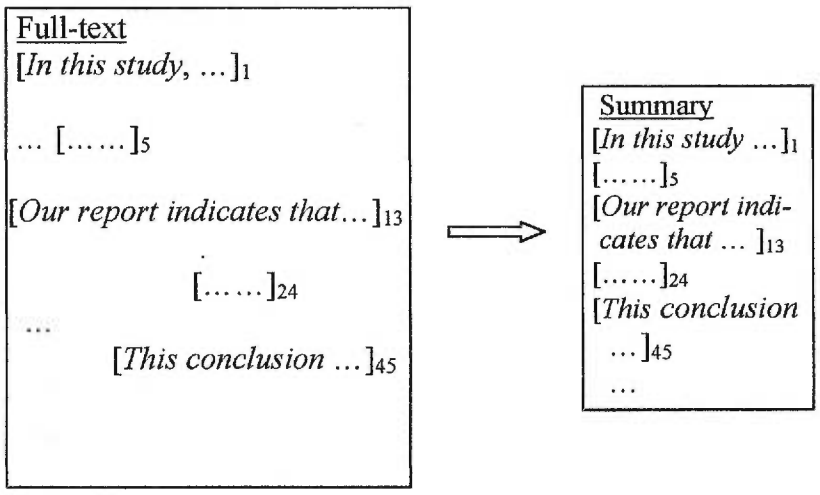


Fig. 2-2: Sentence Extraction by Statistical Technique

2.3.2 Lexical Cohesion Technique

Lexical units in a text cohere in two ways: (a) reiteration; (b) collocation (Halliday & Hasan, 1976). Computation of cohesive strength between lexical units may be based on: (a) number of 'links[12]' between them (Benbrahim & Ahmad, 1995), or (b) the "distance" between them as found in a thesaurus (Morris & Hirst, 1991; Barzilay & Elhadad, 1997). When sentences with the greatest number of links, or from the strongest of chains of related words (see Morris & Hirst,

[12] Consider the sentences: *Apples1 are plenty this year. The apples2 are Quebec-grown.*
For Benbrahim & Ahmad (1995.:327), a 'link' is a connection by repetition between any two linguistic items in separate sentences. In the sentences given above, there is only one link which is between *apples1* and *apples2*. For Halliday & Hasan (1976:329), a 'tie' (besides being a repetition) can include the relation between an element, i.e. *the*, and another, i.e. *apples1* , presupposed by it. Hence, there are two ties between the same two given sentences. This makes a link a more restricted version of a tie.

1991; Barzilay & Elhadad, 1997), are extracted to constitute a summary, the extracted sentences represent text content, even if partially. This second technique in summarization falls between the two extremes of full semantic interpretation (which is expensive to carry out) and frequency count (which ignores any relation that may exist between distinct words). Unlike the statistical technique, the lexical cohesion technique is unaffected by organization of text, or by text genre. We know of only two applications of this technique in summarization. The techniques differ in the algorithm for sentence extraction.

Benbrahim & Ahmad (1995) computed the cohesive strength between sentences based on the number of links between them. A 'bond' is said to exist, if there is an above-average number of links, i.e. the 'bond threshold[13]'. Next, the number of bonds that a sentence has with sentences before and after it is determined. A sentence is then categorized as follows: (a) 'topic opening' - if there are more bonds after than before it; (b) 'topic closing' - if there are more bonds before than after it; (c) 'central' sentence - if there are bonds both before and after it. A summary of desired content, and length is then obtained by selecting the appropriate type and number of sentences: "the user may only wish to look at central sentences, or a combination of central, topic opening, and topic closing sentences" (*ibid.*:330).

Three remarks against this algorithm are: (a) the indeterminacy of bond threshold which is reported to vary within and without a document; (b) the fuzzy distinction between central sentences, and topic opening and topic closing sentences[14]; (c) the methodology is biased toward long[15] sentences which are likely to have more links.

---

[13] The bond threshold is said to be different among texts, and to vary "even within the same text" (Benbrahim & Ahmad, 1995.:328).

[14] We give here some data taken from Benbrahim & Ahmad (1995:332) in support of our argument.
Sentence number, S (number of bonds before **S**, number of bonds after **S**);
TO = topic opening sentence, C = central sentence, TO = topic closing sentence.
e.g. **12** (7,30) is considered as C, whereas **20** (11,35) is TO. Remark: if **12** is C, then **20** should be C.
e.g. **59** (30,3) is considered as C, whereas **79** (33,13) is TC. Remark: if **59** is C, and **79** should be C.

[15] The first TO sentence **4** (2,67) selected in the study by Benbrahim & Ahmad (1995:333) contains 104 words. The sentence is: *The extent to which we are able to make precise and meaningful statements about the nuclear matter distribution and the nuclear charge distribution and the variation in both quantities from one nucleus to another reveals quite clearly the state of our understanding of much more fundamental issues, such as the nature of the interactions between various types of particles and the role of these interactions in scattering phenomena, the subtle balance between various features of the nucleon-nucleon interactions in bound states, and the difference between the average properties of nuclei described by macroscopic models and the specific nuclear structure properties described by microscopic models.*

Below, we describe Barzilay & Elhadad's work which is the better known of the two.

Using Lexical Chains for Text Summarization - Barzilay & Elhadad (1997)

Lexical units in cohesion by virtue of a semantic relation between them, or the fact that they are collocates, result in a 'lexical chain'. Barzilay & Elhadad exploited the reflection of "topic progression" via lexical chains in their algorithm for summarization. Consider the following text (from Barzilay & Elhadad, 1997:12)

> **Mr.** Kenny is the **person** that invented an anaesthetic **machine** which uses **micro-computers** to control the rate at which an anaesthetic is pumped into the blood. Such **machines** are nothing new. But his **device** uses two **micro-computers** to achieve much closer monitoring of the **pump** feeding the anaesthetic into the patient.

To construct a lexical chain, a set of candidate lexical units (given in ***bold***) is first identified, and the relation between them looked up using on-line thesaurus WordNet. For WordNet, a semantic relation can hold between: (a) word forms, e.g. synonymy, or (b) lexicalized concepts, e.g. hyponymy (Miller-G.A, 1998:24).

CANDIDATE WORD	SYNSET(S)[16] AS EXTRACTED FROM WORDNET
Mr.	sense 1: Mister, Mr. (a form of address for a man)
person	sense 1: person, individual, ... (a human being);
	sense 2: person (a person's body);
	sense 3: person (a grammatical category of pronoun ...)
machine	sense 1: machine (any mechanical or electrical device ...);
	sense 3: machine (an efficient person);
microcomputer	sense 1: ... micro-computer (a small computer ...);
device	sense 1: device (an instrumentality invented for ...);
pump	sense 1: pump (a device that ...)

Consider candidate word ***Mr.*** which has only one synset with meaning 'a form of address for a man', abridged to synset $\mathbf{Mr_1}$. Synset $\mathbf{Mr_1}$ may be tied to the next candidate ***person*** via synset $\mathbf{person_1}$ with meaning 'a human being', but not synset $\mathbf{person_3}$ with meaning 'a grammatical category'. We now have a lexical chain of two synsets $\mathbf{Mr_1}$--$\mathbf{person_1}$.

Synset **person$_1$** may in turn be tied to the next candidate ***machine*** via synset **machine$_3$** with meaning 'an efficient person', but not synset **machine$_1$** with meaning 'a device'. The chain is now extended to three synsets: **Mr$_1$**--[17]**person$_1$**==**machine$_3$**. Note that synset **Mr$_1$** may be tied directly to synset **machine$_1$**, without going through the intermediary of synset **person$_1$**. We hence have another possible chain, **Mr$_1$--machine$_3$**. In this way, a set of lexical chains are obtained for a given component[18].

{ **Mr$_1$--person$_1$==machine$_3$, Mr$_1$--machine$_3$** }

Based on the lexical relation between the synsets, an arbitrary number of points[19] is awarded, and the strongest of lexical chain, is selected for a component. Taking another component, say ***device***, the same is repeated and another set of lexical chains obtained.

{ **machine$_1$--micro-computer$_1$--machine$_1$--device$_1$--micro-computer$_1$--pump$_1$,**
machine$_1$==machine$_1$--device$_1$--micro-computer$_1$--pump$_1$, ... }

The strength of all chains are calculated, and the strongest of for each component determined. Length of chain and homogeneity index[20] were found to be good predictors of strength. However, chain members contribute in varying degrees to chain strength. Three heuristics were experimented, and the best was that which selects the sentence containing the first most representative member in the chain, i.e. the most frequently occurring. For example, if the lexical chain

machine$_1$--micro-computer$_1$--...micro-computer$_1$-- pump$_1$

is the strongest for the component ***device***, and **micro-computer$_1$** is the most representative member for the chain, then the sentence containing the first mention of **micro-computer$_1$** will be selected. In this way, sentences are extracted to construct a summary. Only one sentence is

---

[16] A set of words that are interchangeable in some context (WordNet 1.6).
[17] -- : strong chain between synsets; == : extra-strong chain between synonyms and repetitions.
[18] A component was defined as "a list of interpretations that are exclusive of each other." (Barzilay & Elhadad, 1997:13).
[19] Lexical units within the same synonym set (synset), as with repetitions, are considered as "extra-strong", and are given 10 points. If the units are one or more synsets apart, as with hypernyms and holonyms, they are given 4 points. Antonyms which are considered slightly weaker than extra-strong, are given 7 points.

extracted per chain. An advantage of using lexical chains is that concepts represented by several synonyms with low frequency which would otherwise be missed, are brought together. Chains across text segments[21] may be merged, if they have at least one chain member in common.

Thirty texts from popular magazines were tried. The success of the algorithm was said to hinge on a good scoring function for lexical chains. On this matter, we question the arbitrary basis on which points are awarded, as this is consequential on sentence selection. Also, the necessity to build all lexical chains before the strongest of chains may be extracted is a disadvantage. It is claimed that the summary so produced was superior to that carried out by summarizers "in commercial systems such as search systems on the World Wide Web" (*ibid.*:16). However, there was no mention on how their evaluation was carried out. We note here that because only one sentence per semantically distinct chain is extracted, it is improbable the extracted sentences are in any relation with one another – syntactically or even semantically, apart from the fact that they are in the same document, and thus, are likely to be about the same subject. Further, because the first sentence from the strongest of chains was selected, extracted sentences tend to concentrate at the beginning of document, which would mean that important sentences found at the end of document would be missed unless taken into consideration during segmentation of document. Fig. 2-3 illustrates the process.

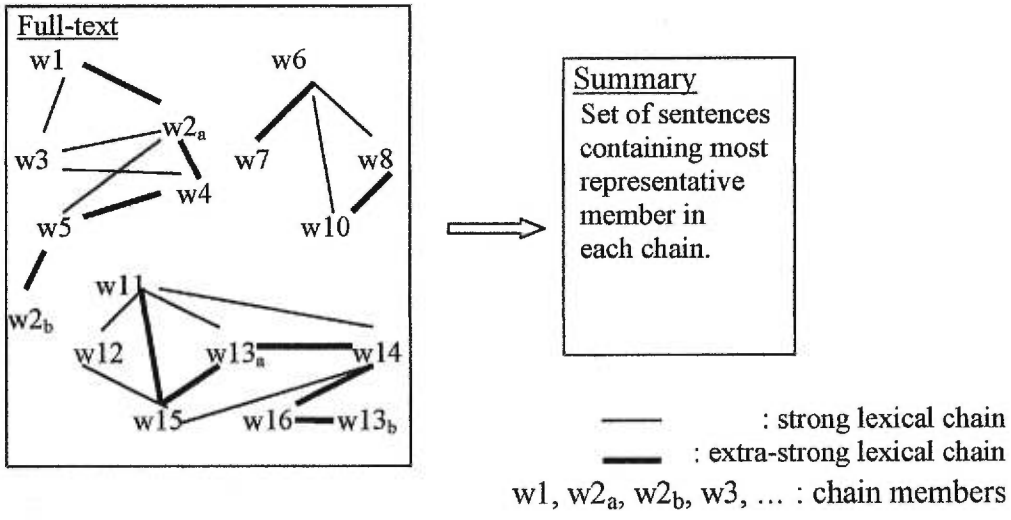


Fig. 2-3: Sentence Extraction by Lexical Cohesion Technique

[20] Homogeneity index = (1 - distinct occurrences/chain length).
[21] A document is segmented before lexical chains are constructed.

2.3.3 Rhetorical Structure Theory (RST) Technique

Rhetorical relations have been studied since the days of Aristotle, and varying sets of relations have been proposed ever since. In their Rhetorical Structure Theory, Mann & Thompson (1987) gave a set of 23 relations[22]. As different rhetorical relations serve different functions, their frequency of occurrence will vary with text genre which serve different purposes[23].

In Mann & Thompson's RST (*ibid.*), a relation is assigned to 'text spans' without recourse to any linguistic means, grammatical or lexical, and hence is ideal for identifying linguistically unsignaled relations. A relation is assigned on the basis of responses to four constraint fields[24]. Because of the nature of the theory, the assignment of RST relations cannot be automated. Note however that current implementations of the theory make use of discourse markers.

The structure resulting from the application of an RST relation is called a 'schema' (see Fig. 2-4). Of the five types of schemas identified, three (i.e. Sequence, Contrast and Joint relations; *ibid.*:73-77) are entirely multinuclear while the other two have a nucleus with either one or two satellites. Instruction manuals tend to have multinuclear schemas. The most common schema is that where one text span, the 'nucleus', is more important than the other, the 'satellite'. Importance is interpreted as "necessary for the interpretation of the text span with which it is in relation"[25]. In RST, there is no maximal constraint on the size of a text span, although the minimal unit is a clause or a nominalization of clause.

Fig. 2-4: Rhetorical Structure Theory Schema

[22] Hovy (1990) has argued that the relations can be reduced to the three basics: Elaboration, Enhancement and Extension.

[23] For example, technical texts are unlikely to have the Motivation relation which has the effect of increasing the reader's desire to perform some action mentioned in an adjacent text fragment.

[24] The first and second constraint fields specify conditions on the text spans of nucleus and satellite respectively, while the third field specifies those on their combination. The fourth field concerns reader's disposition with regards to the text presented in the nucleus, upon reading the satellite, i.e. the effect.

[25] Consider the sentences: (1) She won the lottery last month. (2) The check for a million dollars arrived today. Upon reading (2) the satellite, the reader is more convinced of the truth of (1) the nucleus. The relation between (1) and (2) is one of Evidence.

The application of RST to a text gives a schema of schemas. Satellite text spans may be pruned off the schema until a summary of the desired length is obtained, and without affecting the comprehensibility of nucleus text spans. Marcu (1997:85) showed the nuclei to coincide with important sentences.

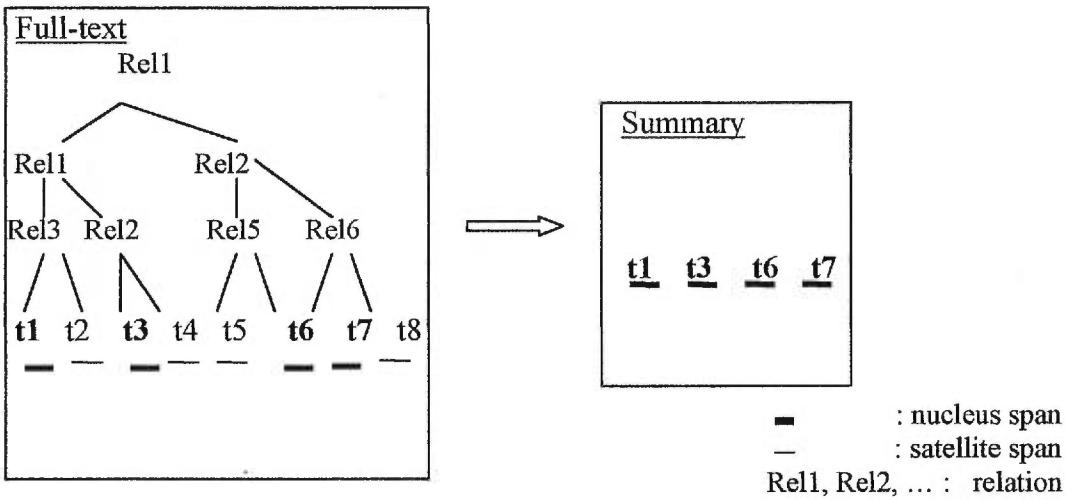
We describe below the work by Ono *et al.* (1994).

Abstract Generation Based on Rhetorical Structure Extraction - Ono *et al.* (1994)

An input text is first subjected to morphological and syntactical analysis. Next, according to the 'connective', or 'sentence predicate'[26] present, a relation is assigned. Where no connective is found, a default relation, i.e. Extension, is assumed. Their set of thirty four relations differs from those proposed by Mann and Thompson (1987). Although RST makes no use of linguistic means in the assignment of discourse relation, Ono *et al.* (1994) did just that without justifying the action taken. The rhetorical structure tree for a paragraph is built up, before that between paragraphs are constructed. To link paragraphs, the connective associated with the first rhetorical relation in the paragraph is used. Once the rhetorical structure tree is fully constructed, demerit points are given. Important sentences receive less demerit points than unimportant sentences, and sentences closest to the root get the least demerit points. Terminal sentences that are more deeply embedded get more demerit points. Through a demerit system, sentences with the least demerit points are selected until an abstract of the desired length is obtained. The sentences extracted constitute the more important sentences in the text. Because of the different levels at which two relations of the same type may occur, it is possible for the relation to be given different demerit points. However, no consideration is given to award different demerit points to different relations, i.e. a Reason relation is no different from an Example relation in terms of demerit points. Of the two types of Japanese expository texts studied - editorial articles from a Japanese newspaper and technical papers, the latter fared better (60%:74%). This was attributed to the presence of more rhetorical devices; editorials were found to be lacking in linguistic clues. A disadvantage of RST is that it necessarily requires the whole text to be subjected to a full analysis in terms of RST relations before satellite spans can be pruned off. The possibility of multiple relation assignment[27] is another disadvantage. An evaluation of RST-produced

[26] E.g. *here ... is described* signals the Direction relation, and *that is to say* signals the Rephrase relation.
[27] Taking the same example in the preceding footnote, the relation can arguably be one of Elaboration.

summaries was made by comparing the agreement in sentences from extracted summaries and those selected by judges. Fig. 2-5 illustrates the process.



Note: The figure was adapted from Fig. 2 in Marcu (1997:86)

Fig. 2-5: Sentence Extraction by Rhetorical Structure Theory Technique

2.4 Some Designations and Processes in Condensation

Reflective of our fragmented understanding of CONDENSATION and its sub-processes is an assortment of designations with or without definition/illustration, and no clear distinction between them.

Now, is condensation the sum of all the processes implicated in summarization? May the operation of displacing a linguistic unit from the end of a sentence or document to the front be considered as condensation? What about the expansion of an acronym or the substitution of a heavy compound or proper name with its acronym?

To better understand what condensation is, and what its constituent processes are, we will examine some designations already proposed or identified by researchers in summarization, and in text generation. For want of a proper definition for condensation in linguistics or in summarization, the definition in general language will be used as basis.

> CONDENSE To make something that is spoken or written shorter, by not giving as much detail or using fewer words to give the same information.
>
> (Longman Dictionary of Contemporary English, (LDOCE), 1995)

In the next six sub-sections, we discuss some designations/processes proposed or identified by various researchers, and in the last sub-section of 2.4.8, we give a categorization of the designations.

2.4.1 Rush, Salvador & Zamora (1971)

In their work on automatic abstracting, Rush, Salvador & Zamora (1971:262-263) discussed how using punctuation as guide, a coordinated or subordinated segment of text in a sentence may be truncated without affecting the rest of the sentence. Sentence S is rendered more concise by truncating or deleting a segment of text. The result is a modified sentence S' which may further be truncated to S" (see Table 2-1).

Table 2-1. Condensation Rules Identified by Rush, Salvador & Zamora (1971)

Rules	Example
1. paraphrase, truncation	(a) S → S': The house was beautiful in the winter~~, but its was more comfortable in the summertime~~$_x$ → [The house was beautiful ~~in the winter~~$_y$ ϕ_x]$_{S'}$ → [The house was beautiful ϕ_y ϕ_y]$_{S''}$;
2. concatenation, embedding	(a) S_1, S_2 → S;
3. fragmentation	(a) S → S_1, S_2;

~~text with a strikethrough~~ = deleted text; S, S_1, S_2, S', S" = sentences;

Except for TRUNCATION, no example is offered for the other transformation types. While CONCATENATING sentences S_1 and S_2 by embedding produces one sentence S, FRAGMENTING S, the inverse of concatenation, gives separate sentences, S_1 and S_2.

2.4.2 Mathis, Rush & Young (1973)

The earliest work we know that was carried out to expressly improve the quality of abstracts constituted of extracted sentences is that by Mathis, Rush & Young (1973). The study proposed five rules to improve the readability of extracted abstracts (see Table 2-2). While Rule 1 and Rule 2 COMBINE sentences by coordination and by subordination respectively, Rule 3 and Rule 5 on GRAPHICAL REFERENCE and CONTEXT MODIFICATION treat "hanging" reference to figure, table, and ordinal numbers. The latter two rules on hanging reference replace the segment of text for which reference to a figure, table or linguistic unit cannot found, with an appropriate string of

words such that the sentence is no longer hanging: *The table presents …* → *A table presents …*, and *The second mechanism is …* → *A mechanism is ….*

Table 2-2. Condensation Rules Identified by Mathis, Rush & Young (1973)

Rule	Example
1. Combination of Sentences by Means of a Coordinate a Conjunction	(a) S_1, S_2 → S_1 coord S_2: The system exceeded the capacity of its present auxiliary equipment + The system was modified for further testing → The system exceeded the capacity of its present auxiliary equipment and was modified for further testing;
2. Combination of Sentences by Means of a Subordinate a Conjunction	(a) S_1, S_2 → S_1, subord S_2: A set of consecutive storage locations is called a memory block + A memory block is labelled by a single word called a codeword → A set of consecutive storage locations is called a memory block, which is labelled by a single word called a codeword;
3. Graphical Reference Rule	(a) Table 2 presents nine areas of endeavor and their associated disciplines → A table presents nine areas of endeavor and their associated disciplines; (b) Figure 2 presents graphically the general model of information transfer → A figure presents graphically the general model of information transfer;
4. Reference Tabulation	(a) If N references are given in document, then generate "N references was given", and if no references, then generate "No references are given".
5. Context Modification	(a) The second mechanism is structural change: … → A mechanism is structural change: …; (b) ~~The second is that~~ reactions to oxygen atoms … → ϕ Reactions to oxygen atoms …;

[+] S, S_1, S_2 = sentences; coord = coordinating conjunction; subord = subordinating conjunction; ~~text with a strikethrough~~ = deleted text; text with thick underscore = text added;

While the former two rules aggregate sentences, the latter two rules which treat hanging references may be considered as rules on substitution. Rule 4 is rather unusual in that it adds on a remark in the abstract with regards to the number of references given in full text. The Reference Tabulation rule computes the number of references given in a document and inserts an

appropriate statement. For example, if N references are given, then a statement *N references are given* is inserted into abstract, and if none, then statement *No references are given* is inserted.

2.4.3 Maybury (1995)

Maybury (1995) suggested ABSTRACTION and AGGREGATION to be distinct types of condensation (*ibid.*:736) sub-processes. While abstraction replaces a series of events with a single event, e.g. the substitution of a number of events "by an overarching event", aggregation factors out the unit (e.g. agent, or patient) in common between them, and consequentially merges events into a single description. While abstraction which is in fact generalization, involves linguistic and/or world knowledge, the process of aggregation requires knowledge on syntax. In his work, Maybury also looked into the use of temporal and spatial lexical anaphors, e.g. *three minutes later* and *four miles west (from here)* in the condensation of text, which our study does not treat.

Table 2-3. Condensation Processes Identified by Maybury (1995)

Condensation Processes	Example
1. Abstraction	(a) movement events + missiles firing + aborted mission → foiled attack event;
2. Aggregation	(a) Site A fired a missile at time t + Site D fired a missile at time t → Site A and Site D were simultaneously fired at time t^{+};

$^{+}$ The example here is a modification of the example given in Maybury.

2.4.4 Sparck Jones (1999)

In her review of the summarization process, Sparck Jones (1999) used the term GENERALIZATION in her definition for a summary (see section 2.1). However, no definition was offered, nor did she say if the term was a substitute for condensation, or itself a condensation sub-process.

2.4.5 Jing & McKeown (2000)

To produce concise texts, Jing & McKeown (2000:178-179) identified six "operations" to edit extracted sentences, which we summarize in Table 2-4. The examples are as given in their article although some segments of text which are marginal to the point that we are making, have been replaced by letters X and Y.

Table 2-4. Rephrasing Operations Identified by Jing & McKeown (2000)

Operation	Example
1. Sentence reduction	(a) ~~When it arrives,~~ $X^+ \rightarrow X$;
2. Sentence combination	(a) X + Y → [X and Y]$_S$;
3. Syntactic transformation	(a) Subject in a sentence is moved from the end to the front.
4. Lexical paraphrasing	(a) point out → note; (b) fits squarely into → hits the head on the nail [sic.];
5. (a) Generalization	(a) a proposed new law that would require Web publishers to obtain parental consent before collecting personal information from children → Legislation to protect children's privacy on-line;
(b) Specification	(b) the White House's top drug official → Gen. Barry R. McCaffrey, the White House's top drug official;
6. Reordering	(a) Place an ending sentence in an article at the beginning of an abstract

$^+$ X, Y = segment of text; [...]$_S$ = sentence;
~~text with a strikethrough~~ = deleted text; text with thick underscore = text added;

Taking the meaning of condensation to be as defined in general language (see the beginning of section 2.4), SYNTACTIC TRANSFORMATION and REORDERING (operations 3 and 6) may not be considered as condensation processes.

SENTENCE REDUCTION which removes "extraneous phrases" is essentially a deletion operation. Corresponding to Maybury's AGGREGATION and ABSTRACTION respectively, are operations SENTENCE COMBINATION: "merg[ing] material from several sentences", and GENERALIZATION: "replac[ing] phrases or clauses with more general descriptions". While Jing & McKeown (*ibid.*:179) defined SPECIFICATION to be "replac[ing] phrases or clauses with more specific descriptions", i.e. the inverse of generalization, the single example given is more precisely an operation which adds on a segment of text to a linguistic unit such that the unit to which it is added becomes informatively more "specific", i.e. an operation that is the inverse of

deletion. Generalizations and specifications are, in fact, substitutions with hypernym and hyponym respectively.

Meanwhile, based on the examples given, what is called LEXICAL PARAPHRASING may, in fact, be seen as substitution with synonymous forms. While the substitute in example (a) is a more concise form, the substitute in example (b) is a longer form. In sub-section 2.4.2, Rush, Salvador & Zamora (1971) gave sentence fragmentation, the inverse of aggregation, as a process in producing concise texts. It would seem then that in summarization, condensation is not just about "shortening" a text, but also, "lengthening" it. Hence, the criterion of "fewer words" is not a critical component to include in the definition of condensation. We restrict the use of the term paraphrase to the reformulation of meaning involving larger units such as clauses and sentences.

2.4.6 Saggion (2000)

In a comparative study of information in abstracts prepared by professionals, and information from particular structural sections of a source document, Saggion (2000:55-63) identified fifteen "transformations" which we summarize in Table 2-5a and Table 2-5b. The examples are as given by Saggion except for some segments of text which have been replaced with letters X, Y or Z.

Table 2-5a. Rephrasing Transformations Identified by Saggion (2000)

Transformation	Example
1. Syntactic verb transformation	(a) ~~Finally we~~ address X^+ → Addresses X'; (b) ~~In this paper we~~ have presented X → Presents X';
2. Lexical verb transformation	(a) ~~It identifies~~ X → X' are discussed; (b) ~~This article details~~ X → X' are described;
3. Verb selection	(a) Running X → The running of X' is described; (b) ~~I~~ define X → Gives an overview of X';
4. Conceptual deletion	(a) ~~Section 2~~ gives X → Gives X'; (b) ~~In this paper, we~~ report X → Reports X';
5. Concept re-expression	(a) ~~We~~ analyse X → Analyzes X'; (b) ~~Our~~ Genie system X → Genie X';
6. Structural deletion	(a) ~~Indeed,~~ X → X';
7. Clause deletion	(a) ~~The work described in this paper addresses these by nothing that~~ X → X'; (b) ~~To emphasize this fact we say that~~ X → X';

$^+$ X, Y, Z = segments of text; X', Y', Z' = unmodified or slightly modified X, Y and Z; $[\ldots]_S$ = sentence; ~~text with a strikethrough~~ = deleted text; text with thick underscore = text added;

Table 2-5b. Rephrasing Transformations Identified by Saggion (2000) (continued)

Transformation	Example
8. Parenthetical deletion	(a) It will show how extending the designer's description of the information processing system ~~(with a language that details how changes within the application occurs)~~ can allow for the construction of applications that are self explanatory → Extending the designer's description of the information processing system can allow for the construction of applications that are self explanatory;
9. Acronym expansion	(a) The work focuses on APIs → The work focuses on application programming interfaces;
10. Abbreviation [sic.]	(a) The future of digital imaging at the National Railway Museum → Discusses the future of digital imaging at the NRM;
11. Merge	(a) Protocol + Address Mapping and Connection Management → Protocol selection, address mapping, and connection management are also described;
12. Split	(a) [This has resulted in a tesseral temporal reasoning system, based on tesseral addressing and using tesseral arithmetic, which offers the advantage that it is directly compatible with existing GIS technology]$_S$ → [A tesseral temporal reasoning system has been designed, based on tesseral addressing and using tesseral arithmetic]$_{S1}$ + [It offers the advantage that is with existing GIS]$_{S2}$;
13. Complex reformulation	(a) [SCULPTOR – an intuitive 3D modeling tool] + [The motivation for our work is to invent a design environment for architects, based on the most recent hard- and software developments]$_{S1}$ + [These are mainly virtual reality (VR) interaction tools, fast graphic libraries, and new approaches in Artificial Intelligence]$_{S2}$ → [SCULPTOR, a 3D intuitive interactive modeling tool, is being developed to create a design environment for architects based on virtual interaction tools, fast graphic libraries, and new approaches in Artificial Intelligence]$_S$;
14. Noun transformation	(a) Business telecommunications prices (UK, Sweden, France, Austria, Germany, …) → business telecommunications prices in Europe; (b) the university of Liverpool → the university of Liverpool, UK; (c) ~~Maxcess Library Systems, Inc. with~~ Maxcess Library System → Maxcess Library System; (d) The 1st experiment → Experiment 1; (e) UK: regulation of cable TV → regulation of cable TV in the UK; (f) Integrating Speech and Natural Language Processing → the integration of speech and natural language processing; (g) The Austrian ~~situation in the field of~~ telecommunication infrastructure → The Austrian telecommunication infrastructure;
15. No transformation	

[+] X, Y, Z = segments of text; X', Y', Z' = unmodified or slightly modified X, Y and Z; $[\ldots]_S$ = sentence; ~~text with a strikethrough~~ = deleted text; text with thick underscore = text added;

Despite the differing designations: SYNTACTIC/LEXICAL VERB TRANSFORMATION, VERB SELECTION, CONCEPTUAL/STRUCTURAL/CLAUSE/PARENTHETICAL DELETION, and CONCEPT RE-EXPRESSION (transformations 1 to 8) are all essentially deletion processes. Except for transformation 8 which deletes parenthetical texts, the linguistic units deleted in transformations 1 to 7 contain either overt presence of author (e.g. *I*, *we*, *our*), or inanimate entities such as *research paper*, article. Saggion refers to such linguistic units as "domain concept". When the presence of author is deleted, an inanimate "domain concept" is personified and used in its place, and if such a personified unit is deleted, we are left simply with *X*.

~~*We*~~ *detail X in article* → ~~*Article details*~~ *X* → *X*

The units deleted in transformations 1 to 7 are essentially linguistic units which concern that part of communication in text which takes place between an author and his reader.

In any formal writing, especially of a scientific or technical nature, no proper name or heavy compound noun may be abbreviated without its acronym having been first made explicit, even if it is a common acronym. When first used, a full form is often followed by its acronym given within parentheses, e.g. *deoxyribonucleic acid* (*DNA*). As repeated mention of a proper name or lengthy compound noun renders a text difficult to read, it is often abridged in repeated mentions, to ensure effective communication of content. In the context of summarization by extraction, any acronym found in selected sentences has to be expanded. As no detail is omitted or added on, but expansion of or substitution with an acronym, the two processes of ACRONYM EXPANSION and ABBREVIATION may not be considered as condensation processes.

While the transformations of COMPLEX REFORMULATION and MERGE aggregate texts, the SPLIT transformation which is the inverse of aggregation, de-aggregates a text.

NOUN TRANSFORMATION is a mixed of processes. Example (a) is more precisely a case of generalization involving world knowledge: UK, Sweden, France, Austria, Germany, etc. are European countries. While a precision (i.e. *UK*) is added on in example (b), in examples (c) and (g) units of text are deleted. As observed by Jing & McKeown (2000) and now Saggion, it would seem that the inverse procession of insertion of a segment of text is part of condensation, even if in a secondary way. Examples (d), (e) and (f) are cases of reformulation of content in different

forms suitable to be used in a sentence, i.e. lexical paraphrasing as proposed by Jing & McKeown (2000). We note that sentences in the abstracts studied by Saggion are often without a subject, e.g. *Addresses ...*, and *Gives an overview of ...*. The abstracts used in his study are from the journals *Library and Information Science Abstracts*, *Information Science Abstracts* and *Computer Abstracts*.

2.4.7 Aggregation and its Typology from Text Generation

In the preceding sub-sections, we saw some designations for condensation from the domain of summarization itself. However, we also have some designations in the name of text combining from text generation. To re-express information from multiple fragments of text more concisely in a single fragment is essentially to condense as redundancies are removed. We discuss the work on sentence combining by Dalianis & Hovy (1993), and give the typology of aggregations surveyed by Reape & Mellish (1999).

2.4.7.1 Aggregation by Dalianis & Hovy (1993)

Dalianis & Hovy (1993:90) used the term AGGREGATION to refer to the "removal of redundancy" during generation. They said that Mann & Moore (1980) were the first to use the term, although Paice (1981) was reported (in Paice, 1990:175) to have coined the same term to mean "the idea of adding adjacent sentences". Dalianis & Hovy (1993:88) proposed eight aggregation rules that fall into four classes: (a) Grouping, (b) Ordering, (c) Casting and (d) Parsimony. We draw the reader's attention to the simplicity of the sentences considered in hypothetical situations, and the complexity of sentences actually constructed in general writing. Not only are sentences found in scientific and technical writing highly complex, but rarely share the same convenient structure necessary for aggregation. The examples given as illustration below are our own.

(a) **Grouping** factors out the redundant common subject, or predicate, element, and in the process, sentences are condensed. We note that the units grouped may be condensed further by substitution with a generic word.

e.g. His Doberman bit the postman. His German shepherd bit the postman.
→ His Doberman and German shepherd bit the postman.

e.g. Monkeys like bananas. Monkeys like papayas.
→ Monkeys like bananas and papayas.

Although Dalianis & Hovy's (*ibid.*) notion of redundancy includes inferable text material, they restrict themselves to the simpler problem of explicitly repeated text material. We give here some examples of inferable text material that should be removed in the production of concise texts.

e.g. Jean likes meat. He likes beef. → Jean likes beef.

e.g. She is not rich. She is poor. → She is not rich. OR She is poor.

e.g. The hunter took a knife. He stabbed the bear.

→ The hunter stabbed the bear.

?The hunter took a knife and stabbed the bear.

(b) **Ordering** concerns the presentation of elements when aggregating text fragments. This makes it part of planning "how to say" in text production. Note that only elements within the "order zone", i.e. those that are in a sequence, and are of the same "rhetorical generality and importance" may be ordered (*ibid.*:97). The ordering priority of elements which was determined to be:

state-change[28] > animate > inanimate > concept-supertype (isa) > attribute > ...

was said to apply "without exception". This is to say that if the following two clauses are aggregated, then the concept element is always presented before the attribute element.

e.g. The pen is red. The pen is a Schaeffer. → The pen is a Schaeffer and is red.

The priority of ordering of elements is believed to be characteristic of most domains. Grouping and Ordering rules are said to be applicable in any order.

[28] State-change, e.g. from idle to busy; animate, e.g. a subscriber; inanimate, e.g. a speech connection; concept-supertype (isa), e.g. is a subscriber, and attribute, e.g. is idle (from Dalianis & Hovy, 1993:97).

(c) **Casting** concerns the consistent use of a verbal construct or a nominal lexical form in use throughout the text. Consequently, it is linked to its presentation, rather than its aggregation. However, repeated use of a verbal construct or nominalized form, provide opportunities for removal of redundancies. Consider the following sentences:

> Grouping factors out the redundant common subject, or predicate, element.
> Ordering concerns the presentation of elements when aggregating text fragments.
> Casting concerns the consistent use of a verbal construct or a nominal lexical form in use throughout the text.
> Parsimony concerns the reduction of verbosity.

The last three sentences which are similarly cast, may be grouped. However, for reasons of readability (see parsimony rule of economy below), they are not. Similarity in construct, nevertheless, contributes to text cohesion. For reasons of variation, casting rule is unlikely to be applied consistently in long texts.

(d) **Parsimony** concerns the reduction of verbosity. Text fragments when compounded upon text fragments, using the classes of rules mentioned above, will inevitably result in highly compacted sentences with low readability. Parsimony rules, i.e. Economy and Repetition, ensure that this does not happen.

The parsimony rule of economy calls for the preference for short sentences to long ones, and where a smaller number of elements is preferred, and if possible with no more than three elements.

e.g. The pen is a red Schaeffer. is preferred to The pen is a Schaeffer and is red.
1 2 1 2 3

The parsimony rule of repetition asks that similar propositional constructs be aggregated. This differs from casting which prescribes the consistent use of a construct or a form. The parsimony rule of repetition is said to apply early during content selection, while that of economy can be applied at any time (*ibid.*:95).

e.g. Eliana is a student at UdeM and is from Chile. + Daniel is a student at UdeM and is from Chile.
→ Eliana and Daniel are students at UdeM and are from Chile.

2.4.7.2 Typology of Aggregation Surveyed by Reape & Mellish (1999)

Reape & Mellish (1999:23-25) surveyed the definitions for aggregation by various researchers, and proposed "broad" and "narrow" definitions for aggregation.

> AGGREGATION (narrow definition): any process which maps one or more structures into another structure which gives rise to text which is more x-aggregated[29] than would otherwise be the case.
>
> AGGREGATION (broad definition): the combination of two or more linguistic structures into a single linguistic structure which contributes to sentence structuring and construction.

We present in see Table 2-6, the typology of aggregations surveyed by them. From the various types of aggregations, we obtained two condensation sub-processes: (a) substitution with a hypernym, e.g. *hummingbird* → *bird*, and (b) simple aggregation itself, e.g. *J is C's sister* + *C is J's brother* → *C and J are brother and sister*.

Table 2-6. Typology of Aggregation Surveyed by Reape & Mellish (1999)

Type	Example	
Conceptual aggregation	*peacock* + *hummingbird*	→ *bird*
Lexical aggregation[30]	(a) Monday(x1), … Friday(x5)	→ *weekdays*;
	(b) Monday(x1), … Friday(x5)	→ weekdays({x1, …x5});
	(c) *more* + *quick*	→ *quicker*;
Semantic aggregation	*J is C's sister* + *C is J's brother*	→ *C and J are brother and sister*
Referential aggregation	*John is here* + *Jane is here*	→ *They are here*
Syntactic aggregation	*John is here* + *Jane is here*	→ *John and Jane are here*
Discourse aggregation	E(nuc(E(nuc(n), (sat(p1))), sat(p2))[+]	→ E(nuc(n), (sat(and(p1,p2)))

[+] E = elaboration relation; nuc = nucleus; sat = satellite.

[29] The nonce term "x-aggregated" was used by Reape & Mellish to show that the definition is not circular.

[30] "Then there are three types of lexical aggregation possible: (a) the mapping of more lexical predicates to fewer lexemes in one step, (b) the mapping of (more) lexical predicates to (fewer) lexical predicates and (c) the mapping of (more) lexemes to (fewer) lexemes." Reape & Mellish (1999:25)

2.4.8 Categorization of Designations Proposed by Various Researchers

Now, to put some order to the designations that are condensation-related, we propose on the basis of the linguistic operation, three groups of sub-processes (see Table 2-7).

(a) GENERALIZATION, if a linguistic unit is replaced with a hypernym;

(b) DELETION, if unit(s) are removed; and

(c) AGGREGATION, when units are combined.

Table 2-7. Overview of Condensation Processes Identified in Previous Research

CONDENSATION TYPE	SUB-PROCESS OF CONDENSATION TYPE (SOURCE[+])	INVERSE SUB-PROCESS OF CONDENSATION TYPE (SOURCE)
GENERALIZATION	Graphical Reference Rule (M,R&Y), Context Modification (M,R&Y), Abstraction (M), Generalization (SJ, J&M), Lexical paraphrasing (J&M), Noun transformation (a) (S), Conceptual aggregation (R&M), Lexical aggregation (a), (b) (R&M)	Specification (J&M),
DELETION	Paraphrase-truncation (R,S&Z), Sentence reduction (J&M), Syntactic verb transformation (S), Lexical verb transformation (S), Verb selection (S), Conceptual deletion (S), Concept re-expression (S), Structural deletion (S), Clause deletion (S), Parenthetical deletion (S), Noun transformation (c), (g) (S),	Noun transformation (b) (S),
AGGREGATION	Concatenation-embedding (R,S&Z), Sentence Combination by Coordinate Conjunction (M,R&Y), Sentence Combination by Subordinate Conjunction (M,R&Y), Aggregation (D&H, M), Sentence combination (J&M), Complex reformulation (S), Merge (S), Semantic aggregation (R&M), Referential aggregation (R&M) , Syntactic aggregation (R&M), Discourse aggregation (R&M)	Fragmentation (R,S&Z), Split (S),

[+] D&H = Dalianis & Hovy (1993); J&M = Jing & McKeown (2000); M = Maybury (1995); M,R&Y = Mathis, Rush & Young (1973); R&M = Reape & Mellish (1999); R,S&Z = Rush, Salvador & Zamora (1971); S = Saggion (2000); SJ = Sparck Jones (1999);

Among the sub-processes identified, a few are inverse operations of the condensation types proposed, which we put in a separate column on the right (see Table 2-7).

Are these all the types constituting condensation, or are there other processes that reformulate content more compactly? In Chapter 5, we will re-work this categorization after considering how language reformulates content concisely, and also, after identifying the mechanisms actually used by authors when writing abstracts for scientific journal articles.

2.5 Fields Potentially Contributing to Abstracting

In the previous section, we saw some condensation processes in abstracting that implicate the use of different lexical units. To this end, any lexical resource or study that is potentially contributing to alternative expressions should be explored. Also, in specialized domains, not only is information expressed in its own characteristic way. Not only do words have their own generic, but also domain restrictions on co-occurrence. Hence, any research into the behavior of language in restricted domains is equally needed. In this respect, we consider potential contributions from two fields: (a) lexicology, and (b) sublanguage.

2.5.1 Lexicology

Great strides made in lexicology have appeared both in their description, and in the tools that embody the description. WordNet, a dictionary-cum-thesaurus project which started in 1985[31], has been available on-line since 1995 (Fellbaum, 1998:xiv). As an on-line source of lexical information, it is already used in various applications, including summarization (see Barzilay & Elhadad, 1997). The Explanatory-Combinatorial Dictionary (ECD) (Melcuk *et al.*, 1995) while not yet available on-line, gives a comprehensive and an exhaustive description of the lexicon[32]. Formally described in 'predicate function'[33], the ECD is an attractive lexical resource for use in any automatic text processing situation, and accordingly, should be investigated. WordNet and

[31] Although the idea was said to have been conceived 20 years earlier (Fellbaum, 1998:xiv).

[32] The French ECD, *Dictionnaire explicatif et combinatoire du français contemporain*, which is available in hardcopy in four volumes (Melcuk, 1984; 1988; 1992; 1999) contains a total of 508 (= 50 + 107 + 171 + 180) entries. A subset of the entries in the French ECD is already in databsse *Dictionnarire de Combinatoire* (DiCo) (see Polguère, 2000 and also www.fas.umontreal.ca/ling/olst).

[33] F(x) = y, where x = keyword, y = its value. For predicate function Synonym, **Syn**(penetrate) = enter.

the ECD may be considered for use in conjunction with techniques in summarization. We discuss them below.

2.5.1.1 WordNet

In WordNet, words are organized based on similarity in meaning within their syntactic category. Synonymous words are said to belong to the same synonym set, i.e. SYNSET (Fellbaum, 1998:xvii). WordNet (v.1.5) contains 60,000 noun synsets (Miller, G.A. 1998:23), 16,428 adjective synsets (Miller, K.J. 1998:47) and 11,500 verb synsets (Fellbaum, 1998:71). Adverbs were added only in 1992 (*ibid.*:xix). A relation can hold between words within a synset; or between words from different synsets. The latter may be part-whole relations, i.e. MERONYMY, and HOLONYMY[34], or specific-generic relations, i.e. HYPONYMY and HYPERNYMY (*ibid.*:37).

When the hypernym of a word, say *monkey*, or *bird*, is looked up, WordNet gives a list of words in increasing hypernymic relation with it (see below). Note that while WordNet currently treats only single words, set phrases for verbs, e.g. *set up*, or *kick the bucket*, and some nouns, e.g. *rule of thumb*, are also encoded.

WORD	HYPERNYMS
monkey	primate < mammal[35] < vertebrate < chordate < animal ...
bird	vertebrate < chordate, < animal ...

In the context of summarization, by substituting words, say *monkey* and *bird*, with the first hypernym, i.e. *vertebrate*, that is in common to both words, sentences may be grouped without introducing unnecessary vagueness. As much as this principled way of selecting lexical units in generalization is most attractive in an automatic summarization situation, it has to be applied with care to avoid producing awkward statements. Consider the resulting aggregated sentence which is rarely articulated in everyday language. Some discretion is required.

Jean has a monkey for a pet. Jean has a bird for a pet.

→ ?Jean has vertebrates /chordates for pets.

[34] E.g. the holonym of *finger* is *hand*.

[35] "primate < mammal" is to be read as "mammal is a hypernym of primate".

Our research will investigate the extent WordNet is adequate as a resource for abstracting journal articles on biology. However, a setback anticipated is how to select the right word sense. Even for a noun as unambiguous as *monkey*, two senses (sense-1 and sense-2) were given in WordNet. While the first hypernym of sense-1 is *primate*, the first hypernym for sense-2 is *child*. As we envisage an automatic summarization situation, the system must not only be able to determine the right syntactic category for a word, but also the right sense among those available.

WordNet has been used by Hirst & St-Onge (1998) to detect malapropism. However, their complaint is that WordNet does not allow chains to be built across syntactic categories. In an earlier work, Morris & Hirst (1991:41-42) remarked that a general thesaurus does not contain lexical relations specific to specialized domains. While this is true, the information presently encoded for the specific-generic relation for words pertaining to animals and plants is impressively rich to be potentially useful in abstracting in a specialized domain such as entomology. However, a quick check on some common terms pertaining to chemistry, e.g. *alkane* and *halogen*, seems to indicate that the domain information included for these words is not as complete as that for biology. *Butane* is not given as a hyponym of *alkane*. Also, while the definition for *halogen* lists *fluorine*, *chlorine*, and *bromine* among its hyponyms, a search for the hypernym of *fluorine*, *chlorine*, and *bromine* does not consistently turn up *halogen* as it should: *halogen* is not given as the hypernym for *fluorine* as is found in the definition of its co-hyponyms.

2.5.1.2 Explanatory-Combinatorial Dictionary

The Explanatory-Combinatorial Dictionary (ECD) has a total of 64 lexical relations[36] (Melcuk, 1996:72), or lexical functions as it is called. Explanatory-Combinatorial Dictionary lexical functions are of two main types: paradigmatic, and syntagmatic. While paradigmatic lexical functions deal with "nomination", syntagmatic lexical functions deal with "combination" (*ibid.*:46).

Among the syntagmatic relations in ECD is **Gener**(ic) which describes the generic word that can syntagmatically follow a word, e.g. **Gener**(republic) = state (see Melcuk, 1996:51). With this, it is meant that *state* can follow *republic* without *republican state* and *republic* being

significantly semantically different. While *republican state* and *condensation process* are acceptable as linguistic units, with *condensation process* and *condensation* being semantically equivalent, we would not normally accept *?banana fruit*, unless in contrast with something similar, say, *banana juice*, even if it is clear that *banana fruit* refers to *banana*. Strictly following ECD guidelines, **Gener**(condensation) = process may be encoded, but not **Gener**(banana) = fruit.

Information encoded in ECD is strictly lexical, i.e. no information pertaining to world knowledge, e.g. part-whole and specific-generic relations such as found in WordNet is encoded. While this is true, this syntagmatic ECD lexical function of **Gener** is, in fact, similar to the WordNet's paradigmatic relation of hypernymy, which ECD does not encode.

To express text content in alternative ways, paradigmatic substitutes, e.g. **Syn**(onym), may be used. The ECD also has lexical functions to describe logical arguments, or deep-syntactic actants as they are referred to in ECD. The lexical function for the first deep-syntactic actant of a verb is $\mathbf{S_1}$, e.g. $\mathbf{S_1}$(parasitize) = {parasitoid, parasite, ...}. Below, we use ECD lexical functions to describe the relation between some words from the sublanguage of biology. To know what the lexical functions signify see Melcuk (1996).

e.g.	$\mathbf{Syn}_{C}$(dissect)	= cut	**Syn**(transfer)	= switch
	$\mathbf{Conv}_{21}$(host)	= parasitize	$\mathbf{S_1}$(parasitize)	= {parasitoid, parasite, ...}
	$\mathbf{S_0}$(parasitize)	= parasitism	$\mathbf{S_2}$(parasitize)	= host
	Cap(hive)	= queen bee	**Equip**(hive)	= {soldier bee, worker bee, drone}

The utility of syntagmatic relations although immediately less obvious for condensing extracted sentences, is anticipated to be useful in situations where one is generating from conceptual units. Syntagmatic relations which are absent in WordNet, link morphologically distinct but related lexical units, e.g. **Magn**(feed) = actively, intensively.

[36] Far more than the eight or so in WordNet.

The ECD is intended to describe lexical relations in natural language. But because their values in restricted domains may be different, an ECD for specialized language is welcomed. Consider

General language		Biology	
Mult(egg)	= clutch	**Mult**(egg)	= egg mass

The effort required to encode the information for a given lexical unit for entry into the ECD is overwhelming, and with its present *modus operandi*, we are pessimistic about its completion early enough to be of any immediate utility. A solution would be for partial automation of the encoding process using a dictionary editor (Melcuk *et al.*, 1995:207), and for the more important lexical functions, only then may its utility be tested. A system, DiCo (*Dictionnaire de Combinatoire*), to automate data entry for a subset of lexical functions is in development for the French ECD (Polguère, 2000).

2.5.2 Sublanguage[37]

A sublanguage is that part of a language which can be described by a specialized grammar (Sager, 1982:9). In sublanguage texts, text segments that are predictable to a domain reader are "unnecessary", and are often omitted. For example, in the domain of cookery, determiners are often left out, e.g. *Put chicken in oven* instead of *Put the chicken in the oven.* In biology, the relative pronoun, e.g. *that*, and auxiliary verb are omitted wherever expected to be predictable to a (domain) reader. Also, commonly omitted is the preposition *of.* In the following examples, we have re-inserted the lexical units that were omitted from full text sentences. These re-inserted lexical units are given in **bold** within square brackets.

e.g. *E. harmandi* was found aboard spiders **[that were]** associated with the forest floor.
All of the flowers offered 1 µl **[of]** scented 30% sucrose solution …
We provide evidence **[that suggest]** that undertaking specialists are … of the same age.

[37] For a short introduction see Kittredge & Lehrberger (1982: 1).

The omission of implicit lexical units may take place during abstracting as seen in the following example, ft-sentence → ab-sentence.

e.g. these "ergatoid" males … monopolize all young queens which are reared in their maternal nests over several weeks.
→ They … increase their share in copulations with the virgin queens reared in their nests.

Besides such abridged constructs, special word usage uncommon or even unacceptable in general language are normal for a given sublanguage. For example, while we may board buses and planes, and construct houses and factories in general language, in scientific language an insect can board another life form, and larvae can construct cocoons.

e.g. Mantispids board spiders.
The larva constructs a pupal cocoon within the spider egg sac.

Halliday (1993:71) discussed some features of scientific English which make processing difficult. These features include: (a) interlocking definitions (e.g. *radius ~ diameter*), (b) technical taxonomies (e.g. *climate ~ temperature*), (c) special expressions (e.g. *solving the open sentence over D*), (d) lexical density (i.e. the concentration of lexical words is high), (e) syntactic ambiguity (e.g. *lung cancer death rates*), (f) grammatical metaphor[38] (e.g. not *how quick cracks in glass grow*, but *glass crack growth rate*), and (g) semantic discontinuity (i.e. knowledge is required to link *cleaner factories* with *anti-pollution laws* in *strong anti-pollution laws … have resulted in cleaner factories*). In a text production situation such as abstracting, knowledge of characteristic constructs and particular usage of words different from that in general language which bear on the production of an abstract with a sublanguage tone, should accordingly be noted and exploited.

The regularity with which words in a sublanguage co-occur has been exploited by Hirshman & Sager (1982:28). Using the idea of what is called an INFORMATION FORMAT, i.e. a table-like structure whose columns correspond to the word classes of basic sentence types of the sublanguage, they were able to automate the retrieval of information encoded within sublanguage structures which is not possible with general language. Information format

[38] "substitution of one grammatical class, or one grammatical structure by another" Halliday (1993:79).

organizes sublanguage sentence types into a compact tabular representation so that the document content can be quickly inspected. Taking an example from their study on clinical reporting, in a basic sentence sequence of subject-verb-object where the verb is of a given verb-type (e.g. *have, develop*), the subject (e.g. *patient*) and object (e.g. *cough, fever*) will be from particular word classes (i.e. PATIENT and SIGN-SYMPTOM respectively).

This research on sublanguage is pertinent to abstracting as it tells how information in restricted domain may be extracted to instantiate pre-determined templates. Paice & Jones (1993) exploited this information format in abstracting using constructs signaled by indicator phrases in his work using the template approach to summarization.

2.6 Concluding Remarks

Research on summarization has mostly stopped at content selection. While there is some work on condensation, the designations proposed are rather mixed. Based on whether a linguistic unit is replaced, deleted from, or combined with another unit, we categorized the designations proposed or identified by various researchers into three main groups: (a) GENERALIZATION, (b) DELETION and (c) AGGREGATION.

Inverses of these processes were also identified. During condensation, linguistic units may also be expatiated by: (a) substituting with a more specific unit, which Jing & McKeown (2000) intend by "specification"; (b) adding on of linguistic unit(s) to incorporate more details into the existing unit which Saggion (2000) refers to under "noun transformation", and (c) the de-aggregation of information in a sentence into multiple sentences which Rush, Salvador & Zamora (1971) refer to as "fragmentation", and which Saggion (2000) refers to as "split".

Because of the ambiguity in meaning of "specification" between that of replacement with a more specific unit proposed by Jing & McKeown (2000), and the meaning of "detailed instruction about how something should be designed or made" (LDOCE, 1995), we propose provisionally using PARTICULARIZE as the inverse of GENERALIZE. Particularize is given as the antonym for generalize in WordNet. While DELETION removes linguistic units, we propose using INSERTION as the inverse process, in the sense of the adding on of linguistic units as used in

linguistics. Because AGGREGATION is already well-accepted to mean the combining of text segments, we propose using DE-AGGREGATION as the inverse operation.

As seen from the processes identified by various researchers, condensation in summarization is not just "not giving as much detail or using fewer words to give the same information" (see general definition given at the beginning of section 2.4), but also includes the adding on of units of information. Provisionally, we amend the definition in general language to that given as follows for content condensation in the context of summarization.

> CONDENSATION The process of making something that is spoken or written shorter, by not giving as much detail or using fewer words to give the same information, or by augmenting information such that a unit is informatively more compact, or more explicitly expressed.

And, we propose its constituent processes to be the following three categories:

(a) GENERALIZATION,

(b) DELETION, and

(c) AGGREGATION.

Closely linked to these groups of condensation processes are their inverse operations which seem to play an important, although secondary role in abstracting. When we re-work our typology of condensation processes in Chapter 5, we will also discuss their inverse processes.

Chapter 3

Methodology

Chapter 3 describes the methodology in the study of content condensation mechanisms in abstracting (aim 1 of the present research). A description of the study corpus is given in section 3.1, and the preparation of the documents for study in section 3.2. In section 3.3, we discuss some examples of matches between full text (ft-) sentences and abstract (ab-) sentences. Ft-sentence(s) that best match an ab-sentence are assumed to have been selected for abstracting by an author. These ft-sentences are henceforth referred to as SELECTED FT-SENTENCES. We end the chapter with a discussion on some difficulties in the matching process. Some statistics on the distribution of selected ft-sentences are reported in the next chapter.

3.1 Corpus

Fifty-seven articles downloaded from two journals, *Behavioral Ecology and Sociobiology* (bes) and *Oecologia* (oec) (Springer Publications) and divided into four sub-corpora (bes1, bes2, oec1 and oec2) of 13-15 articles each constitute the study corpus (see Appendix III for complete titles of articles). The documents have the basic sections of Abstract (A), Introduction (I), Method (M), Results (R) and Discussion (D).

The abstracts of the articles studied are written by the author-researchers themselves. While Saggion & Lapalme (1998:73) consider professional abstracts to be "better structured in content and form because they are produced from the reading of the document following specific strategies", we argue that "following specific strategies" does not mean that important content has been selected, but simply that the guidelines of the abstracting service have been adhered to. Also, who other than an individual author himself knows best which content to extract for abstracting, or how to organize and word an abstract? Compared to a professional abstractor, the author of a document is for us the ultimate if not the sole authority with regards to both content and structure of abstracts. As one author[39] of the articles studied so aptly put it (pers. comm.):

> "To the best of my knowledge, abstracts of all the articles in biological journals ... are prepared by the authors of the articles. ... the authors are in the best position to interpret

[39] Prof. K.V. Yeargan from the Dept. of Entomology, Univ. of Kentucky.

and summarize the key points of their research (i.e., better than the journal editor, better than a third party, etc.)."

3.2 Preparation of Text for Study

For identification purposes, all sentences in full text and abstract were given a code to indicate its location in document. For example, a sentence with location code [R-2-1] is the first sentence in the second paragraph of the Results section. An alphanumeric code, such as bes2-9638145 (journal-year_volume_page), gives the source of document in journal. In the study, proper names, e.g. *the Cincinnati Nature Center (Clermont Co., Ohio)*, and scientific names, e.g. *Schizocosa ocreata (Hentz)* are considered as single words. All words found in headings, captions, or within parentheses were excluded from word count, but not from consideration in the identification of full text sentences used in abstracting, i.e. as sources of information in abstract sentences.

Some statistics on corpus: (a) their size, in terms of number of sentences and number of words, and (b) distribution, are as given in Table 3-1.

Table 3-1. Statistics on Corpus

	full-text (ft)	abstract (ab)	reduction factor
Corpus size	7938 sn; 175,613 wd	534 sn; 11,975 wd	15:1 (sn or wd)
Range of size of article	62–269 sn; 1,552–6,333 wd	5–21 sn; 109–415 wd	7:1–31:1 (sn or wd)
Av. size of article	139 sn; 3,081 wd	9 sn; 210 wd	15:1 (sn or wd)
Range of sn length	4–129 wd	7–80 wd	
Av. sn length	22.12 wd	22.42 wd	
Distribution of sn length	≤ 9 wd: 5%	≤ 9 wd: 4%	
	10–14 wd: 16%	10–14 wd: 14%	
	15–19 wd: 23%	15–19 wd: 23%	
	20–24 wd: 20%	20–24 wd: 22%	
	25–29 wd: 16%	25–29 wd: 17%	
	30–34 wd: 9%	30–34 wd: 10%	
	35–39 wd: 5%	35–39 wd: 6%	
	40–44 wd: 3%	40–44 wd: 3%	
	≥ 45 wd: 2%	≥ 45 wd: 2%	

total no. documents = 57; sn = sentence; wd = word;
reduction factor = no. ft-sn (or wd)/ no. ab-sn (or wd);

3.3 Identification and Selection of Ft-sentences Used in Abstracting

On the basis of identity in word, stem, and semantic similarity at the level of word and expression, between a ft- and an ab-sentence, a manual search was made for sentences most likely to have been used as the basis of abstracting[40].

First, for each content[41] or lexical word in an ab-sentence, all words in full text that are identical or share the same stem are underscored with a thick line. Next, after narrowing down on some ft-sentences with a high number of words underscored, i.e. CANDIDATE FT-SENTENCES, ft-words/ft-expressions that are synonyms or equivalents for the other lexical units in the ab-sentence, are underscored with a thin line. This exercise helps to identify in an organized way linguistic units in full text and in abstract that correspond to each other, and to reduce the degree of subjectivity in identifying sentences used in abstracting, which is inevitable in any manual selection of ft-sentences, particularly when multiple processes come into play.

Sentence selection was guided by length of verbatim match, number of words with same stem, number of synonymous words and most DIRECT MATCH, i.e. a match which requires lesser number of manipulations to get from a linguistic unit in full text to its corresponding linguistic unit in abstract. A ft-sentence may be considered as a match even if there are modifiers in an ab-sentence that are unaccounted for, but not if the words in abstract are nouns or verbs. The reason is because modifiers which communicate marginal content may be omitted during abstracting, but not nouns and verbs which communicate core content. To help a reader identify corresponding units, some units are subscripted. Multiple-word linguistic units may be underscored with a dotted line to show that they form a unit. These marks do not affect the selection process.

The proportion of lexical units in ft-sentence which do not correspond to any lexical unit in abstract, is not a factor in selection. When two sentences have about the same number of units underscored, the most direct match is preferred. The set of ft-sentences selected may be assumed to have been the set extracted by a summarization technique among those described in the previous chapter.

---

[40] The present author is a trained entomologist.
[41] Content words are words that have stateable lexical meanings (Crystal, 1997).

A continuum of matches of varying difficult are encountered. While direct matches may be a simple extract-and-prune, or splice-and-join, complicated matches require multiple manipulations and various kinds of knowledge. In the first three sub-sections that follow, we describe some simple more common types of ft-ab sentence matches identified: (a) one-ft-one-ab, (b) two-ft-one-ab, and (c) one-ft-two-ab, and in section 3.4, we illustrate some problematic matches.

3.3.1 One-ft-one-ab Sentence Match

Consider Example 3-1. Only one clear candidate ft-sentence was found. The parenthetical text is deleted. *Slave-added colonies* is substituted with *treatment colonies* which is a domain synonym. Linguistic unit *had* is replaced with a unit *produced* that is more explicit in its meaning. For both operations, experimental knowledge is required to effect substitutions.

Example 3-1

Full text (bes2-9638145)	Abstract
T7#: When the number of sexual offspring$_z$ was adjusted for colony size, slave-added colonies$_x$ had$_y$ significantly more sexuals$_z$ than the controls ~~(one-tailed test, as the only expected change after adding food or slaves is an increase of sexuals)~~. [R-2-3]	A7: When colony size was adjusted to the number of sexual offspring$_z$, the treatment colonies$_x$ produced$_y$ significantly more sexual offspring$_z$ than the controls. [A-1-7]

\# = ft-sentence that is a partial match for ab-sentence;
~~strikethrough~~ = text not used in writing ab-sentence;

In Example 3-2, two candidate sentences were identified for [A-1-5]. However, based on our criterion of longest verbatim match, [R-1-1] which is the better candidate is assumed to have been selected and used in abstracting. Selected candidates are marked with a hash sign #, to differentiate it from unselected candidates which are unmarked. In Example 3-1 and Example 3-2, only one ft-candidate was used in abstracting. However, as many as four or more ft-sentences may be used to write an ab-sentence.

Example 3-2

Full text (bes2-9638145)	Abstract
T5: Because *F. podzolica* slaves are active foragers, we expected that the increase in $slaves_x$ would enhance the food harvest of the colony, particularly the following spring when sexual offspring are developing in the nest, yielding more sexual $offspring_x$ in the $colonies_y$ with $slaves\ added_y$ than in the control $colonies_z$. [I-3-4] T5#: The proportion of slaves was significantly $higher_x$ in the treatment $colonies_y$ than in the $controls_z$. [R-1-1]	A5: The proportion of slaves was significantly $higher_x$ in the slave-added $colonies_y$ than in the control $colonies_z$. [A-1-5]

= ft-sentence that is a partial match for ab-sentence;

3.3.2 Two-ft-one-ab Sentence Match

Consider Example 3-3. While [I-2-1] is a good candidate, it provides only part of the information to write [A-1-1]. The rest of the information in ab-sentence has its source in [I-1-3]. Unlike preceding examples, in Example 3-3 the information for the ab-sentence comes from two ft-sentences. A conjunction was used here to combine the segments of text.

Example 3-3

Full text (bes2-9638145)	Abstract
T1b#: Facultative slavemakers$_x$ are able to forage, nurse their brood ~~and construct their nest like free-living ants, and hence~~ colonies without slaves$_y$ are common$_z$. [I-1-3] T1a#: *Formica subnuda* is a facultative slave-making ant$_x$, ~~and belongs to the *F. sanguinea* group~~. [I-2-1]	A1a: *Formica subnuda* is a facultative slave-making ant$_x$, A1b: and colonies without slaves$_y$ are often found$_z$. [A-1-1]

= ft-sentence that is a partial match for ab-sentence;
~~strikethrough~~ = text not used in writing ab-sentence;

In Example 3-4, as in Example 3-3, the information for the ab-sentence has its source in two ft-sentences.

Example 3-4

Full text (oec1-97109265)	Abstract
T6#: Gender affected$_x$ biomass gained, food consumed, RGR$_y$ and ECI$_z$, but had no significant effect on duration of stadium and the relative consumption rates. [R-5-1] T6#: In general, the females consumed more$_w$, grew more$_w$, grew faster$_w$, and converted ingested$_z$ matter to biomass more efficiently$_w$ than did males. [R-5-2]	A6: There were several effects$_x$ of gender: biomass gained, food consumed, relative growth rate$_y$ and efficiency of conversion of ingested food$_z$ to biomass were higher$_w$ for females than males. [A-1-6]

= ft-sentence that is a partial match for ab-sentence;

Sentences in scientific texts are often highly complex. In Example 3-5, we break ab-sentence [A-1-3] into two smaller fragments of A3a and A3b. For ab-fragment A3a, there are two possible candidate ft-sentences, [R-5-1] and [D-1-3]. As the former ft-sentence does not contain all the information in ab-fragment A3a, and the latter does, [D-1-3] is considered the better candidate. The information for ab-fragment A3b is found in [D-1-4]. Two ft-sentences were used to write ab-sentence [A-1-3].

Example 3-5

Full text (oec2-97109454)	Abstract
T3a: Larval survival was greater$_y$ on unoccupied than on ant-occupied acacias. [R-5-1] T3a#: Although larval density and survival are higher$_y$ on acacias not occupied by ants, shelters serve as a partial refuge from the ant *P.ferruginea*. [D-1-3] T3b#: Shelters provide *Polyhymno* larvae access$_y$ to an otherwise unattainable and poorly defended host plant$_x$. [D-1-4]	A3a: Although larval density and larval survival are higher$_y$ on acacias not occupied by ants, shelters serve as a partial refuge from the ant *Pseudomyrmex ferruginea* (Hymenoptera: Formicidae), A3b: which defends *A.cornigera* plants$_x$; thus, shelters provide *Polyhymno* larvae access$_y$ to an ant-defended host plant$_x$. [A-1-3]

= ft-sentence that is a partial match for ab-sentence;

In the ab-sentence [A-1-3], the source of information *(Hymenoptera: Formicidae)* is not found. In the corpus studied, it is standard practice to add information on the Order and Family to which an insect belongs during abstracting, regardless of whether this information was given or not in the document.

3.3.3 One-ft-two-ab Sentence Match

In the preceding examples, we saw that one or more ft-sentences may be the source of information for an ab-sentence. However, the converse may also occur, i.e. the information in an ft-sentence may be split or reformulated in separate ab-sentences.

In Example 3-6, the information in [M-5-2] was split between two ab-sentences, [A-1-5] and [A-1-6]. There were few sentences that were de-aggregated in the study.

Example 3-6

Full text (oec2-97117258)	Abstract
T5: In small Kentucky streams, green sunfish$_z$ are one of the most potentially dangerous fish predators. [M-3-2] T5#: Sunfish$_z$ attack water striders from below in deeper water, [M-5-2a]	A5: Green sunfish$_z$ occupy stream pools and attack water striders from below. [A-1-5]
T6#: while fishing spiders perch$_x$ vertically on rocks and overhanging vegetation along the shore where they may catch and lift$_y$ water striders off the water's surface. [M-5-2b]	A6: In contrast, fishing spiders hunt$_y$ along stream shorelines where they perch$_x$ on overhanging vegetation or rocks and attack$_y$ water striders near shore. [A-1-6]

= ft-sentence that is a partial match for ab-sentence;

Where there are multiple sources for the same information in an ab-sentence, precedence in document is not a criterion. On the basis of verbatim match, [D-3-5] is the better source over ft-sentences [R-6-2] and [R-6-3] combined in Example 3-7. However, in Example 3-8, ft-sentences [R-4-1] and [R-5-1] combined is considered the better source over [D-15-2].

Example 3-7

Full text (oec1-97109265)	Abstract
T7: At the warmer thermal regime, both males and females consumed less$_x$ at 2 and 3 mmol of tomatine. [R-6-2] T7: At the cooler thermal regime, the amount that males consumed was unaffected$_x$ by the levels of tomatine consumed by their prey, whereas the females showed a sharp increase$_x$ in consumption at the lowest level of tomatine in the diet of their prey. [R-6-3] T7#: The tomatine by thermal regime by gender interactions for biomass gained and food consumed are difficult to interpret but indicate that temperature and diet are affecting$_x$ males and females differently. [D-3-5]	A7: Furthermore, the effects$_x$ of thermal regime and tomatine on food consumption and biomass gained differed for females and males. [A-1-7]

= ft-sentence that is a partial match for ab-sentence;

Example 3-8

Full text (oec2-97109313)	Abstract
T7#: There were no significant differences$_x$ in spider species richness or diversity between control and ant-free trees$_y$ on any sample date. [R-4-1] T7#: The exclusion of ants$_y$ did not have a significant effect on the overall spider community structure. [R-5-1] T7: In addition, the significant increase in the absolute abundance of hunting spiders was not strong enough to significantly alter the spider community structure. [D-15-2]	A7: Spider diversity and community structure did not differ significantly$_x$ between control and ant-removal trees$_y$. [A-1-7]

= ft-sentence that is a partial match for ab-sentence;

3.4 Some Difficulties in Sentence Matching

While we conducted the matching process as procedurally as we could to simulate a matching algorithm, the task was not always as direct as suggested in the preceding sub-sections. The difficulty may be attributed to various factors, such as repetition of the same information, interplay of condensation sub-processes, different sources of cognitive knowledge implicated at various points, etc., all of which need investigation (in future work, but not in the present research reported). We illustrate with some examples below.

3.4.1 Repeated Information

In Example 3-3, we saw an ab-sentence constituted of text segments spliced off two candidate ft-sentences. In the example, there were no other close candidates. In Example 3-9, not only were the sources not as clear-cut, but the information which was reported in the Results section was re-stated and repeated within the Discussion section.

Example 3-9

Full text (oec1-97109265)	Abstract
T5#: At the cooler thermal regime$_x$ (21:10°C), the stinkbugs$_z$ gained less weight and took almost twice as long to complete the stadium, which led to lower relative consumption and growth rates than at the warmer thermal regime$_y$ (26:15°C). [R-1-2] T5#: At the cooler thermal regime$_x$, duration of stadium was almost twice as long as at the warmer thermal regime$_y$. [R-8-1] T5#:Chlorogenic acid-fed prey prolonged the stadium of the stinkbugs$_z$ at the cooler thermal regime$_x$. [R-8-2] T5#: In contrast, chlorogenic acid had a negative effect$_w$ at the spring thermal regime$_x$ on stadium duration and relative consumption rate but not at the warmer, summer thermal regime$_y$. [D-1-7] T5#: Similar to the effects$_w$ of chlorogenic acid on stinkbugs$_z$, *M. sexta* caterpillars were most affected$_w$ by rutin and chlorogenic acid at a cool thermal regime$_x$. [D-1-8] T5#: At the summer thermal regime$_y$, tomatine had a negative effect$_w$ on biomass gained and relative growth rate; at the spring regime$_x$, chlorogenic acid prolonged stadium duration. [D-7-3]	A5: At the cooler thermal regime$_x$, stadium duration was prolonged when the predators$_z$ were given chlorogenic acid-fed prey, but at the warmer thermal regime$_y$ there was no such effect$_w$. [A-1-5] [R-1-2] + [R-8-1] + [R-8-2] → [A-1-5]

= ft-sentence that is a partial match for ab-sentence;

Three possible sets of solution were found: (a) ft-sentences [R-1-2], [R-8-1] and [R-8-2]; (b) ft-sentences [D-1-7] and [D-1-8], and (c) ft-sentence [D-7-3], for ab-sentence [A-1-5]. Where there are multiple solutions, a choice has to be made to determine which gives the most direct match. Overall, fewer manipulations are required with solution (a) than with solution (c). As

knowledge is a factor in determining which sentence was used in abstracting, the matching process was unavoidably subjective. More subjectivity is involved if domain, or world knowledge is implicated compared to linguistic knowledge. Meanwhile, experimental knowledge is not accessible to a non-author-researcher. While solution (a) may require mechanical manipulation and linguistic knowledge, solutions (b) and (c) require world knowledge to get from *spring regime* to *cooler thermal regime*, and from *summer thermal regime* to *warmer thermal regime*. Based on our criterion of verbatim match, solution (a) was chosen to be the best match.

Consider Example 3-10, where the same information is found in three consecutive sentences within the same section. While all three sentences are equally good candidates, the third sentence was selected, on the basis of closest match, to be the ft-sentence used in abstracting.

Example 3-10

Full text (bes1-9638253)	Abstract
T6@: ~~There was also a significant difference in the responses of female spiders to the two video stimulus types;~~ fewer$_x$ females were receptive to the asymmetric male video. [R-3-1] T6@: ~~In the experiment with two independent groups of females;~~ fewer$_x$ females showed receptivity to the asymmetric male video than to the control (symmetric) male video. [R-3-2] T6@: ~~The paired-design video experiment showed similar results:~~ female receptivity was lower$_x$ with the asymmetric male video stimulus. [R-3-3] !! repeated info;	A6: Female receptivity to the asymmetric video image was lower$_x$. [A-1-5] [R-3-3] → [A-1-6] [R-3-1] ≅ [R-3-2] ≅ [R-3-3]

@ = ft-sentence that is a full match for ab-sentence; !! = notes/comments;
~~strikethrough~~ = text not used in writing ab-sentence;

3.4.2 Multiple Sources

In the sub-section 3.4.1, we saw the possibility of having different sets of solutions for an ab-sentence. Unlike the previous sub-section, in Example 3-11 we do not have clear-cut sets of solution for writing an ab-sentence. Combinations of two, three or four sentences from the four candidates, i.e. [I-7-1], [I-7-3], [I-7-4] and [D-7-3], could have been the source of information for ab-sentence [A-1-4].

Example 3-11

Full text (oec2-97109454)	Abstract
T4#: *P. ferruginea* ants act as the primary herbivore defense of *A. cornigera* plants. [I-7-1] T4#: Most nonant acacias possess$_x$ allelochemicals and tough leaves that deter or prevent herbivory by insects. [I-7-3] T4: *A. cornigera* plants do not possess$_x$ allelochemicals and have relatively tender leaves. [I-7-4] T4#: *A. cornigera* plants have poor$_x$ chemical and mechanical defenses, and may be particularly vulnerable to herbivores which avoid ant defense. [D-7-3]	A4: *P. ferruginea* ants act as the primary antiherbivore defense of *A. cornigera* plants, which lack$_x$ the chemical and mechanical defenses of non-ant-defended acacias. [A-1-4] [I-7-1] + [I-7-3] + [D-7-3] → [A-1-4]

= ft-sentence that is a partial match for ab-sentence;

While ft-sentences [I-7-1] and [D-7-3] are strong candidates, ft-sentences [I-7-3] and [I-7-4] are weak candidates. Ft-sentence [I-7-3] was included as a candidate because of lexical units *nonant* and *acacias* which corresponded to units in *non-ant-defended acacias* in the end segment of ab-sentence [A-1-4]. While ft-sentence [I-7-3] was included as it is more direct to get from *nonant acacias* than which avoid *ant defense* to *non-ant-defended acacias*, its inclusion is non-critical as the information is already in ft-sentence [D-7-3]. Part of the information in lexical unit *allelochemicals* is shared by lexical unit *chemical*. Also, lexical unit *possess* in ft-sentence [I-7-3] which is lexically linked to *lack* by an antonym relation, provides a more direct link than unit *have poor* in ft-sentence [D-7-3]. Ft-sentence [I-7-4] was not considered as a candidate because it contains information which is already found in the other candidate sentences. Three ft-sentences were used in writing ab-sentence [A-1-4].

3.4.3 Dispersed Sources

In Example 3-12, the similarity between units from full text and abstract is less obvious than in the preceding examples. One can arguably say that ab-sentence [A-1-10] has it basis in one or two of the selected ft-sentences, instead of three. However, because each ft-sentence on its own is a weak candidate, all three ft-sentences were chosen as combined candidates.

Example 3-12

Full text (oec2-98117133)	Abstract
T10#: Our results provided partial support for$_x$ the hypothesis that atmospheric nitrogen deposition produces altitudinal$_y$ patterns in leaf phytochemistry with consequences for herbivores. [D-2-1] T10#: Our foliar nitrogen data for paper birch from Mt. Moosilauke were consistent across 2 years and matched the results of Lang *et al.*, indicating$_x$ that this spatial$_y$ pattern is consistent across time. [D-2-3] T10#: Nitrogen deposition might explain$_x$ altitudinal$_y$ patterns in *L. dispar* defoliation ~~that have been previously attributed to soil moisture gradient~~. [D-7-8]	A10: Atmospheric nitrogen deposition offers a promising hypothesis to explain$_x$ and predict$_x$ some important spatial$_y$ patterns in herbivory. [A-1-10] [D-2-1] + [D-2-3] + [D-7-8] → [A-1-10]

= ft-sentence that is a partial match for ab-sentence;
~~strikethrough~~ = text not used in writing ab-sentence;

Consider the operations required for the transformations in (a) and (b).

(a) provided partial support for the hypothesis + were consistent ... indicating + might explain
→ offers a promising hypothesis to explain and predict

(b) altitudinal patterns in leaf phytochemistry with consequences for herbivores + spatial pattern + altitudinal patterns in *L. dispar* defoliation
→ spatial patterns in herbivory

In transformation (a), we need to recognize that *provided partial support for* and *might explain* are equivalent to *offers a promising hypothesis to explain and predict*. To substitute were *consistent ... indicating* with *to explain and predict*, experimental knowledge is required. For the transformation in (b), while the link between *herbivores* and *herbivory* is direct, domain and

experimental knowledge is required to link *defoliation* with *herbivory*. Also, knowledge is required to know that *altitudinal patterns* may be considered as a kind of *spatial patterns*.

3.4.4 Domain and Experimental Knowledge

In Example 3-13, there are few lexical units in ft- and ab-sentences that share stems. Domain knowledge is required to know that when one talks of the *genetic relatedness* between X and Y, one is talking about the genetic distance between X and Y, and when one says that the *genetic distance is low* between X and Y, we mean that X and Y are *close relatives*. While around and *less than* may be approximately equivalent semantically, experimental knowledge is required to be able to reformulate *around 4 h* more specifically as *less than 4 h*.

Example 3-13

Full text (bes2-9842009)	Abstract
T5#: ~~Finally,~~ the effect on host discrimination of the genetic relatedness$_x$ between the female marking the egg and the female detecting the mark was studied. [I-6-4] T5#: ~~The acceptance of parasitized hosts was linked to the oviposition experience of the females, since females in experiment 2, which had oviposited eight times before the test, accepted only 13.3% of the parasitized hosts in their first three encounters and females tested again after varying amounts of time (experiment 3, T2)~~ also rejected parasitized hosts early in the sequence$_y$. [R-1-4] T5#: ~~Hence,~~ learning is ~~slower when the genetic distance is greater, and~~ faster when the genetic distance is low$_x$ (females encountering their own marks) or when females were already experienced ~~(time T2)~~. [D-3-8] T5#: The time necessary to forget completely$_z$ is around 4 h in absence of stimuli. [D-1-6] !! good example of need for "restoration" of omitted implicit units to abstract;	A5: Learning lasted$_z$ less than 4 h and occurred earlier in a series$_y$ when the female marking the egg and the one detecting the mark were close relatives$_x$. [A-1-5] [I-6-4] + [R-1-4] + [D-3-8] + [D-1-6] → [A-1-5]

= ft-sentence that is a partial match for ab-sentence; !! = notes/comments;
~~strikethrough~~ = text not used in writing ab-sentence;

Meanwhile, a complicated mix of knowledge is required to "restore" implicit text segment *what has been learned* which is missing from ft-sentence [D-1-6], before text segment *The time necessary to forget completely* can be linked to text segment *Learning lasted.* In a full text, there is space to develop a sentence to an abridged sentence with implicit units omitted. However, if such an abridged sentence happens to be selected, the implicit units may have to be restored. When restored, ft-sentence [D-1-6] reads as:

> The time necessary to forget completely what has been learned is around 4 h in absence of stimuli.

3.4.5 Cognitive Knowledge

In Example 3-14, reasoning is required to get from the two propositions of *host discrimination has to be learned,* and *host discrimination can be forgotten* in ft-sentence [I-4-5] to text segment *Learning is generally predicted not to be important in host discrimination* in ab-sentence [A-1-1]. Because host discrimination has to be learned, and because host discrimination can be forgotten, it is implied that learning which is acquired through host discrimination cannot be important. Linguistic transformation and manipulation are not sufficient to produce text segment A1a. Domain knowledge and reasoning is required.

Example 3-14

Full text (bes2-9842009)	Abstract
T1a#: ~~Van Lenteren and Bakker, van Lenteren and Klomp et al. supposed that~~ host discrimination has to be learned, ~~because females that had not shown oviposition experience for some time behaved in the same way as inexperienced females, indicating that~~ host discrimination can be forgotten~~,~~. [I-4-5]	A1a: Learning is generally predicted not to be important~~,~~ in host discrimination
T1b#: In parasitoids, learning occurs more frequently in the host-habitat location process than in the host acceptance process because the cues associated with this foraging stage should be less variable. [D-6-1]	A1b: by parasitoids, because the stimuli involved are less variable than those used in habitat location. [A-1-1]

= ft-sentence that is a partial match for ab-sentence;
~~strikethrough~~ = text not used in writing ab-sentence;

3.4.6 Conditional Statement

In Example 3-15, the information for ab-sentence [A-1-2] may be found in ft-sentences [I-2-2], [I-3-1] and [I-3-3]. However, strictly speaking the ab-sentence may not be written on the basis of these ft-sentences alone. The propositions in the ft-sentences do not report findings, but hypothesize on what might occur. The hypothesis needs to be confirmed, and ft-sentence [D-7-6] provides the confirmation, even if it does not provide the expression, except for the last segment.

Example 3-15

Full text (bes1-9639061)	Abstract
T2#: ~~Under these circumstances, parasitized caterpillars represent poorer quality oviposition sites than unparasitized individuals do, and, like many other parasitoids,~~ *V. canescens*$_w$ is capable of distinguishing between these different host-types when deciding whether or not to lay$_x$. [I-2-2] T2#: ~~Given that the chemical marker which enables *V. canescens* to detect prior parasitism is sufficiently potent that it enables foraging wasps to identify their own and other's progeny some time after initial attack,~~ additional ovipositions$_x$ are not thought to occur in error$_y$. [I-3-1] T2#: ~~However, since this behaviour places immature wasps into competition with each other,~~ deliberate superparasitism can only be adaptive$_y$ for solitary species$_w$ when there is some probability that second-laid progeny can win the battle for host-possession. [I-3-3] T2: ~~However, it is further argued that~~ this behaviour can only prove adaptive$_y$ for solitary species$_w$ providing that an offspring laid$_x$ into a parasitized host has some chance of survival. [D-1-4] T2#: ~~This supports the view that~~ *V. canescens*$_w$ adopts an adaptive pattern$_y$ of differential superparasitism, which reflects the chances that her offspring can survive in any particular superparasitized host. [D-7-6]	A2: For solitary species$_w$, the decision$_y$ to lay additional eggs$_x$ should therefore be based on$_y$ the probability of superparasite survival in any superparasitized host. [A-1-2] [I-2-2] + [I-3-1] + [I-3-3] + [D-7-6] → [A-1-2] [D-1-4] ≅ [D-7-6]

= ft-sentence that is a partial match for ab-sentence;
~~strikethrough~~ = text not used in writing ab-sentence;

3.5 Concluding Remarks

While the matching process in the study was most of the time not too problematic, knowledge is needed. As it is a tedious task to carry out this matching process which is an essential part of the study, a semi-automatic procedure is welcomed to this end to encourage more similar studies on the same or different corpus. While a semi-automatic matching environment cannot decide on which ft-sentence to select, such an algorithm can alleviate a researcher's task by picking out candidate sentences to present to him for manual selection.

From the ft-sentences selected for abstracting, the linguistic units retained in the final abstract, and the way the units are presented, we know what is considered to be important by an author himself from what was included and what was omitted. However, we remind that in an actual semi-abstracting situation, these decisions on what to include or leave out from an abstract, and how to present the content, i.e. generation-related issues, have to be decided by an abstractor, or decided for a semi-automatic system.

Some data from the study will be discussed in the next chapter.

Chapter 4

Some Statistics on Sentences Used in Abstracting

In the previous chapter, we discussed the methodology for identifying ft-sentences used in abstracting. In section 4.1, we present some statistics on the source and number of the ft-sentences selected. Although the distribution and reduction factors over sections do not tell us what the condensation mechanisms are, the findings throw light on: (a) the factor in reduction during content selection and during content condensation; (b) the different reduction factors for each section, which is indicative of the relative importance of a given section, and the possible exploitation of text structure in abstracting structured scientific documents; and (c) the importance of aggregation in content condensation by sentence combining.

Our findings show that while about two-thirds of ab-sentences were written using two or more selected ft-sentences, the average length of ft- and ab-sentences were not significantly different. The implication here is that content in selected sentences must have been condensed to give the abstract. The total number of selected sentences was reduced by half during abstracting.

As mentioned in Chapter 1, one part of our research seeks to obtain some statistics on abstracting for the corpus studied. Using data obtained from ft-sentences determined to have been used in abstracting by an author himself, we present in section 4.2 our case on why some features which are commonly used as indicators of important content, are not reliable as basis for sentence selection. We however feel that the indicators may and should be used in deciding between the more likely of sentences to have been used in abstracting.

4.1 Some Statistics on Selected Ft-sentences

4.1.1 Distribution over Sections

While there are differences among documents on the proportion of ft-sentences from each section selected for abstracting, the tendency is clear: sentences selected for abstracting by an author himself are mostly from the Introduction section, and least from the Method section. The peak of the graphs in Fig. 4-1 give the proportion of documents and the percentage of sentences selected from different sections for abstracting. Consider Fig. 4-1a and Fig. 4-1b. For about one-third of corpus (of 57 documents), about half of the sentences selected for abstracting came from the Introduction section, none from the Method section (for exact figures see Table A4-1 in Appendix IV).

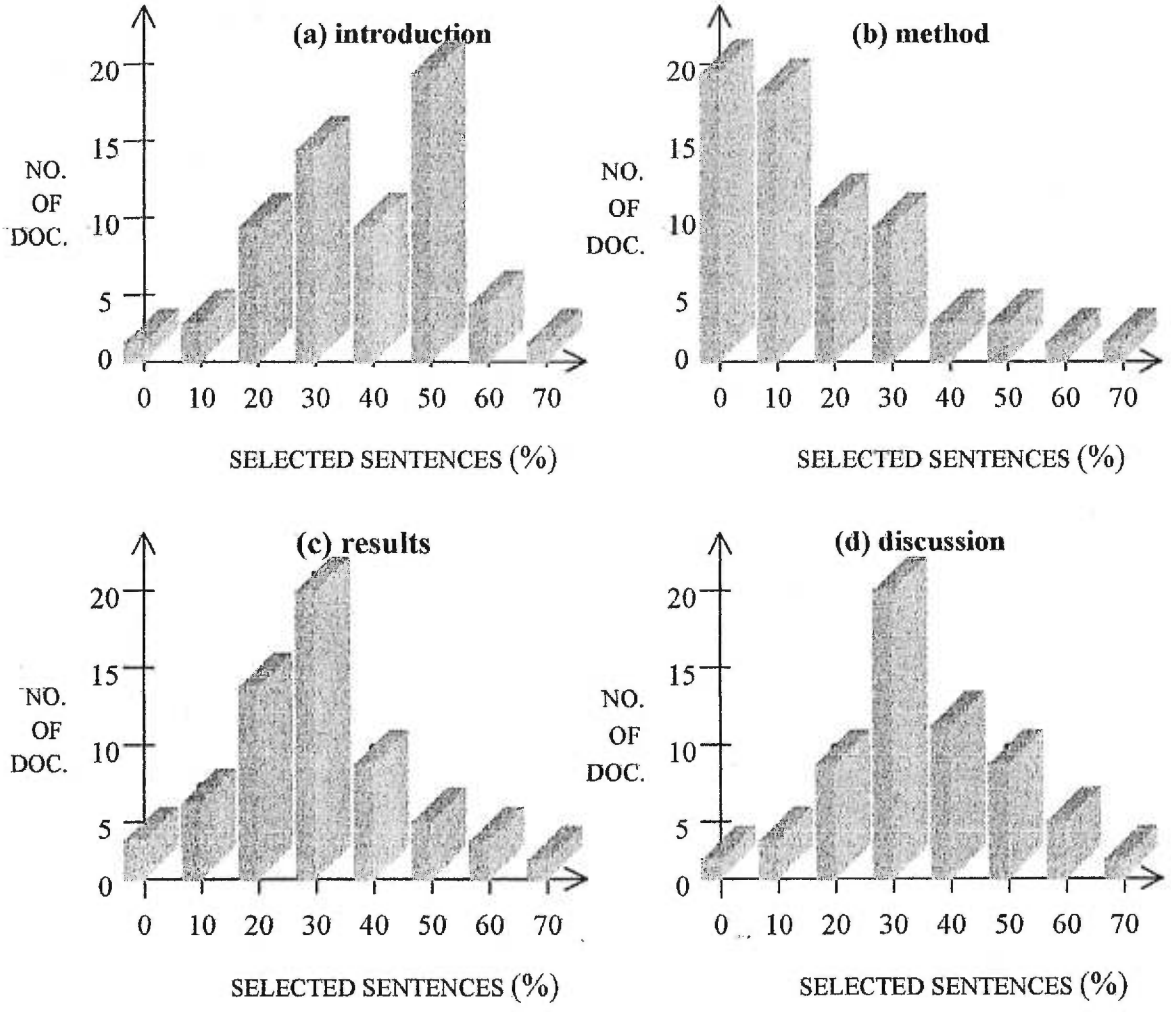


Fig. 4-1. Distribution of % Sentences Selected for Abstracting per Document

On average about 17 sentences are selected per document for abstracting in the proportion of roughly 6: 2: 4: 5 (Introduction: Method: Results: Discussion). The lowest number of sentences selected per document for abstracting is 8, and the highest, 26. In our corpus of articles on biology with author-prepared abstracts, about 34% of sentences selected come from Introduction and 31%, from Discussion (see Fig. 4-2).

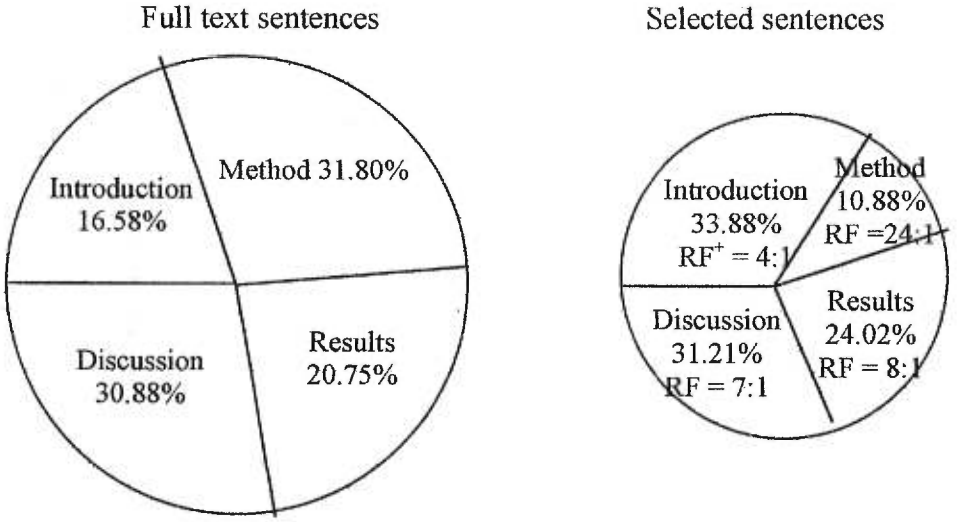


Fig. 4-2. Distribution of Sentences: (a) in Corpus, and (b) Selected for Abstracting
(RF^+ = reduction factor; for exact figures, see Table A4-2 in Appendix IV)

For their mixed *Computer Science* corpus which did not have the prototypical structure of Introduction-Method-Results-Discussion, Saggion & Lapalme (1998:77), found that 40% of the information came from the Introduction section. Their study which was carried out on professional abstracts focussed on "structural parts of the parent document" such as title, sub-headings, captions and first and last sections. It is only when the information is not found in these locations, that the other parts of the full text are considered.

4.1.2 Reduction Factor

While the overall reduction factor from document to abstract is about 15: 1 (total number of all ft-sentences: total number of all ab-sentences = 7938:534), the reduction factor by section is wide-ranging, being greatest for Method (24:1) and lowest for Introduction (4:1) (see Fig. 4-2 above). While the amount of text dedicated to a section may not be taken to be indicative of its importance because of the varying reduction factors depending on each section, there is general consensus in the proportion of sentences to select for abstracting. This observation is a reflection

of the location of important sentences, and hence, the potential of exploiting distribution of sentences by section, indirectly text structure, for abstracting. This feature if used necessarily applies to structured documents, such as journal articles. In what Kupiec *et al.* (1995) refer to as "paragraph feature", they were in fact exploiting text structure when their algorithm considered certain number of paragraphs from the beginning and end of a document which roughly correspond to the sections of Introduction and Discussion.

The total number of sentences in abstracts (i.e. 534) is about half of the total number of sentences selected for abstracting (i.e. 974). This means that to get to the final abstract, selected sentences were condensed by a factor of 2:1. It is the condensation mechanisms in getting from the selected sentences to the abstract which is the focus of our research. In the next section, we take advantage of the statistics indirectly available from our study, to verify the reliability of three features commonly used in extraction algorithms, namely (a) sentence length cut-off feature, (b) fixed-phrase feature, and (c) paragraph feature.

4.2 A Case against some Features Used in Content Selection

Statistical techniques which are the most explored of summarization by extraction approach, use a mix of text cues to select important sentences to constitute an abstract. Despite claims of success, we are not convinced of the reliability of some of the features used. The meaning of a given text is fundamentally in the content words constituting it and in their combination. However, text cues, location, etc. may be used to point to their importance. While we cannot propose any working algorithm for abstracting, we will consider in turn each of the three features of: (a) sentence length cut-off, (b) fixed-phrase, and (c) paragraph, which Kupiec *et al.* (*ibid.*:69) said gave the best results in their study, and argue our case.

4.2.1 Sentence Length Cut-off Feature

Kupiec *et al.* (*ibid.*) asked that sentences shorter than a given threshold, say five words, not be extracted for abstracting. In our corpus, only 0.40% of sentences were five words or shorter in length (see Table A4-3 in Appendix IV). While Kupiec *et al.* (*ibid.*) did not give the percentage of short sentences in their study corpus of engineering-related articles, we do not expect the

number to be significant. In a short survey of four articles each from two different domains[42], none of the sentences (470 and 393 respectively) contain five or less words.

Based on the insignificance of the proportion of extremely short sentences, we question the effectiveness of including this feature in an extraction algorithm. If lexical density is a feature in abstracting algorithms, extremely short sentences with few content or lexical words (see Table 4-1) would automatically be excluded.

Table 4-1. Four and Five-Word Sentences from Corpus

Four words	Five words
Larvae pupated inside shelters. [I-8-7; oec2-97109454]	Eggs hatch in late spring. [I-4-2; oec1-99120268]
Two experiments were performed. [M-7-4; oec2-98117420]	This paper addresses that question. [I-5-7; oec1-99120274]
We conducted two G-tests. [M-6-3; oec1-99120268]	Workers were not individually marked. [M-4-5; bes2-9639293]
Spiders were monitored daily. [M-4-5; oec1- 97111570]	The following parameters were recorded. [M-2-1; oec1-98115154]
Adults emerge in June. [I-3-2; oec1-98115154]	The balloons were never eaten. [R-1-9; bes1-9945161]
Selection was primarily directional. [R-4-7; bes1-9945161]	All colonies produced sexual offspring. [R-2-1; bes2-9638145]
This was not found. [R-9-3; bes2-9741151]	This difference was statistically significant. [R-2-9; oec1-99120274]
	Reasons for this are not known. [D-2-3; oec2-98118381]
	This result has two implications. [D-6-4; bes2-9945047]

Had the sentence-length feature in Kupiec *et al.*'s (*ibid.*) algorithm been used to exclude short titles and captions from being used in abstracting, less than 2% of ab-sentences in our corpus had its source in headings and captions. And so, we do not find this feature to be pertinent for sentence selection in the scientific corpus investigated by the present research. While we did not verify the use of headings and captions for abstracting in other types of corpus, we note that Saggion (2000:49) in his study of computer science-related articles, found 23% of the

[42] *Information & Management* (average size = 120 sentences) and *Neurobiology of Disease* (average size = 98 sentences).

information to be from subtitles and captions. However, unlike our study which considered the whole document, Saggion (*ibid.*) focussed on first and last sections, captions and titles, and it is only when the information is not found that the other parts of a document is checked. This might explain the high percentage found. While we worked with a softcopy of the document, Saggion (*ibid.*) used a hardcopy.

4.2.2 Fixed-phrase Feature

It is commonly believed that sentences accompanied by text cues or phrase indicators, such as *In this study*, *The results suggest*, *In conclusion*, etc, are good candidates for abstracting. In our corpus about four out of five (1124/1366)[43] of sentences cued with any of the following lexemes of CONCLUDE, INDICATE, RESULT, SHOW, STUDY, SUGGEST, which would have been considered likely candidates by current summarization techniques were not selected by an author himself for abstracting (see below, and Fig. A4-1 in Appendix IV for more examples).

> In this study, we examine the impact of regenerative asymmetry on male mating success and fighting ability. [I-2-6; bes1-9638253]
>
> The results suggest a potential for FA measurement error of +55% to –27%, respectively, less than many published values. [R-1-3; bes1-9945087]
>
> In conclusion, allelochemicals can negatively affect both herbivores and their insect predators, but the effects depend on temperature. [D-8-1; oec1-99120252]

This dispels the popular belief that cues such as these are reliable indicators of importance. We remind that an author himself is for us the authority on what is, or is not important for abstracting.

Further, while not all cued sentences are important, not all important sentences are cued. About three out of four (732/974) sentences selected for abstracting did not contain any of the lexemes of CONCLUDE, INDICATE, RESULT, SHOW, STUDY and SUGGEST, which are likely to be found in indicator phrases used by current techniques to locate important sentences. Hence, some sentences which are good candidates for abstracting are missed out. While we admit that a feature is not used on its own, but in combination with other features, it is contributive to the

---

[43] For exact figures see Table A4-4 in Appendix IV.

weight of a sentence. As the eventual selection of a sentence is based on its weight, this brings into question the claims of success of extraction algorithms using this feature. While we have no access to the full list of indicators used by Kupiec *et al.* (1995:69), the two lexemes of CONCLUDE and RESULT are found in the 26 "fixed-phrases" used.

4.2.3 Paragraph Feature

In the paragraph feature, Kupiec *et al.* (*ibid.*) considered only the first ten and last five paragraphs of a document. If this consideration of specific numbers of paragraphs is intended to correspond to the Introduction and Discussion sections, then the paragraph feature is promising: about two-thirds of selected sentences in our corpus came from these two sections, although a significant quarter (109/458) came from the Results section (see Table 4-2). Authors of the documents in our corpus appeared to have exploited text structure in the presentation of important content. In our corpus, the Introduction and Discussion sections averaged five and eight paragraphs respectively.

Table 4-2. Initial and Final-Position Sentences Selected for Abstracting

Sub-corpus	Introduction		Method		Results		Discussion	
	No. Init+Fin	No. Para.	No. Init+Fin	No. Para.	No. Init+Fin	No. Para.	No. Init+Fin	No. Para.
bes1	37	73	13	117	27	99	27	100
bes2	36	59	13	101	19	69	29	130
oec1	50	70	14	100	30	69	39	97
oec2	38	66	12	141	33	106	41	137
total	161	258	52	459	109	343	136	464
probability[+]	0.31		0.06		0.16		0.15	

No. Init+Fin = number of initial/final positions sentences that were used by an author in abstracting;
No. Para. = number of paragraphs; probability = No. Init+Fin / (No. Para. x 2);
[+] the figure should be slightly higher as quite a few paragraphs in document had only one sentence.

The beginnings and endings of paragraphs are said to receive more emphasis than others (Alley, 1996:67). If this is true, then initial and final-position of paragraphs are strategic locations to place important sentences. Kupiec *et al.* (1995) refined the paragraph feature by taking the location of a sentence within a paragraph into consideration. Findings from our study support this refinement to the paragraph feature. Of the 974 sentences determined to have used in

abstracting by an author, 406 selected sentences in the Introduction, Results and Discussion sections were in initial/final position. From the 1065 paragraphs (258 + 343 + 464) from these three sections in our corpus, we have a total number of 2130 possible initial/final position sentences. The probability of an initial/final position sentence from any of these three sections being an important sentence is about 0.19 (406/2130) (see Table 4-2). This probability should be marginally higher as some paragraphs contain only one sentence. We excluded the Method section as it is the least likely section to find important sentences for abstracting. The use of location in text, specifically first sentence of a paragraph has been used in sentence extraction (see Baxendale, 1958; Seuren 1998). The probability of an initial/final position sentence cued with any of the following lexemes: CONCLUDE, INDICATE, RESULT, SHOW, STUDY, SUGGEST, to be a selected sentence for abstracting is 0.25 (116/471) (see Table A4-5 in Appendix IV).

4.3 Concluding Remarks

The data from our study on abstracting in a scientific corpus with structured documents showed there to be clear consensus in the disproportionate selection of sentences from different sections which is reflective of their unequal importance in abstracting. The greatest proportion of sentences selected by an author came from the Introduction and Discussion sections respectively, and least from the Method section. On average about 17 sentences per document were selected from our corpus in the proportion of 6:2:4:5 (Introduction: Method: Results: Discussion). It would then seem then that text structure may be exploited as a potential feature in sentence extraction by section for the corpus type studied.

While Kupiec *et al.* did not explicitly mention the exploitation of text structure in their algorithm, they did so when they considered fixed number of paragraphs from the beginning and the end of a document, which we assume to correspond to the Introduction and Discussion sections respectively. However, our figures appear to be the converse of their figures which was for a different corpus type. Our study showed the Results section to be no less important compared to the Introduction and Discussion sections. About 25% of sentences came from the Results section. Overall, nine out of ten of sentences selected by an author for abstracting came from these three sections.

On the basis of section, and first and last sentence in paragraph, one in five sentences extracted in our corpus was a sentence selected by an author himself. Location of sentence within paragraph appears to be a justified refinement to text structure in this type of text. Accordingly, text structure and location are confirmed to be important in selecting sentences for abstracting in texts on biology, and should be further explored.

While cues and sentence length are common features in current sentence extraction algorithms, our studies showed these two features to be unreliable. Four out of five sentences that were cued were not selected for abstracting, and three out of four sentences that were selected for abstracting were not cued. Meanwhile, because of the low occurrence of short sentences of five words or less, we do not consider the inclusion of such features to be of any significance.

While cues or fixed phrases may be unreliable as basis of sentence selection, they remain nevertheless indicators of importance. And, because of the ease with which such indicators may be exploited, we propose that such cues be used to choose between the more likely of candidate sentences for selection in abstracting, although not as basis in sentence selection. Of the 406 initial/final-position sentences from the Introduction, Results and Discussion sections that were used in abstracting by an author, 116 selected sentences were cued, i.e. with any of the following lexemes of CONCLUDE, INDICATE, RESULT, SHOW, STUDY and/or SUGGEST. The probability of an initial or final position sentence being selected for abstracting is higher when it is cued.

Contrary to some studies (see Saggion, 2000), our study did not show captions and titles to be important sources of information to select for abstracting.

Chapter 5

Condensation Sub-processes

In section 2.4, we categorized the processes proposed/identified by various researchers in summarization and text generation into three main groups of condensation sub-processes, namely generalization, deletion, and aggregation. However, the question we ask is if there are other types of condensation processes in abstracting.

In section 5.1, we discuss some ways in which the English language reformulates content concisely. In section 5.2, we determine indirectly via a comparative study the condensation sub-processes applied by authors when abstracting. On the basis of what we know from these two discussions we revise the provisional categorization obtained earlier for a more comprehensive typology of condensation in abstracting in particular, and summarization in general.

Because inverse operations appear to accompany condensation processes, inverse condensation processes will also be discussed, and their definitions provided, as with condensation processes themselves.

5.1 Content Reformulation in Condensed Form

Below we discuss some ways in which the English language reformulates content in condensed form: (a) GENERALIZATION; (b) NOMINALIZATION and (c) COMPOUND WORD FORMATION.

5.1.1 Generalization

To generalize is "to put a principle, statement, or rule into a more general form so that it covers a larger number of examples" (LDOCE, 1995). By generalizing over lexical units, the essential meaning of a text is communicated, although less precisely. While no reduction in text is implicated in one-for-one replacements.

e.g. His Doberman bit the postman. → His dog bit the postman.

e.g. Monkeys like bananas. → Monkeys like fruits.

Where multiple units of text are involved, the length of a text is reduced.

e.g. His Doberman bit the postman. His German shepherd bit the postman.
→ His dogs bit the postman.

e.g. Monkeys like bananas. Monkeys like papayas. → Monkeys like fruits.

If generalization is applied during summarization, semantically related documents will be linked as related terms are brought together via a superordinate. And should a summary be used in a search instead of keywords, accuracy in retrieval will be affected. A searcher succeeds in retrieving relevant documents not known to him (i.e. higher recall[44]). However, a superordinate item will also lower precision as more documents are retrieved. For this reason, we reiterate here the importance of choice of superordinate items to use in a summary. Superordinate items should be no more hypernymic than necessary, to avoid introducing unnecessary vagueness. To say that *his animals bit the postman* is to introduce imprecision. The unit replaced (i.e. *animals*) can refer to any animal. Also, to say that *monkeys like food*, raises questions as to why such a statement was made. Is it unusual for monkeys to like food?

e.g. His Doberman bit the postman. His German shepherd bit the postman.
→ His animals bit the postman.

e.g. Monkeys like bananas. Monkeys like papayas. → Monkeys like food.

More, we caution against applying generalization in a strict way, as it can result in dubious statements. Consider the example below. It is not true that all animals like bananas.

e.g. Monkeys like bananas. Birds like bananas. → ?Animals like bananas.

[44] Precision is the proportion of relevant documents retrieved to the total no. of documents retrieved (relevant and irrelevant). Hence, if x = no. relevant documents retrieved, z = no. irrelevant document retrieved, then, precision = (x / (x + z)) (Cleveland & Cleveland, 1990: 149). Recall is the proportion of relevant documents retrieved to the total no. of relevant documents (retrieved and unretrieved).

Not only is knowledge other than linguistic knowledge required to select the appropriate hypernym for the situation, but decisions have also to be made with regards to what to generalize and to what level. Indiscriminate application of generalization wherever possible can produce uninformative sentences by stating the obvious.

e.g. Monkeys like bananas. Birds like papayas. → Animals like food.

In section 2.4.7.1, we saw that sentences may be aggregated by grouping, which factors out common elements. While grouping does not have the advantage of bringing together related documents, nor lead to a significant reduction in length of text, it does not run the risk of making dubious statements.

Generalization is a simple case of substitution with a hypernym. If the substitute is a hyponym, we refer to the process as particularization as suggested in section 2.6, or simply substitution, if the substitute is a synonym or a synonymous expression. However, if segments of text are aggregated during generalization, then we consider the process under aggregation.

5.1.2 Nominalization

Crystal (1997) defined NOMINALIZATION as the formation of a noun from another word class, or the derivation of a noun phrase from an underlying clause. For our purpose, the discussion on nominalization will be restricted to the formation of nominalized form(s) from an underlying clause.

Consider the hypothetical proposition "N_1 $V_{transitive}$ N_2", where N_1 and N_2 are respectively the agent and object of a transitive verb, $V_{transitive}$. S_{action} and S_{result} are substantive forms of the verb. A nominalized element is often qualified by other lexical unit(s) from the proposition (see Table 5-1). Nominalization not only compacts information, but at the same time, draws attention to an element within the proposition, e.g. an argument of the verb, action, or the proposition itself, the result. The stem of the element focalized is retained.

Table 5-1. Nominalization

ELEMENT FOCALIZED	NOMINALIZATION RULE	EXAMPLE[+]
Agent	$N_1 + V + N_2$ → N_2 + V-er, → N_2–V-ing + N_1	*Mantispids board spiders.* → *spider boarder,* *spider-boarding mantispid*
Object	$N_1 + V + N_2$ → (N_1–)V-ed + N_2	*Mantispids board spiders.* → *(mantispid-)boarded spider*
Action	$N_1 + V + N_2$ → N_2 + S_{action} (+ hypernym of S_{action})	*Mantispids board spiders.* → *spider boarding* *Larvae enter the book lungs.* → *book lung-entering (behavior)*
Result	$N_1 + V + N_2$ → (N_1 +) N_2 + S_{result}(V) → N_1 S_{result}(V) of N_2	*Pilots report bird strikes.* → *(pilots') bird-strike report* → *pilots' report of bird strikes*

[+] The examples given here are taken from texts used in a preliminary study.

Further attention may be drawn to the element by syntactically placing it in subject position. Drawing attention to a linguistic unit is more easily carried out with nominalized forms than with a clause. With nominalized forms, an author is able to communicate his message efficiently, yet free to manipulate the text to convey his intention. In a study of grammatical subjects in scientific discourse, Vande Kopple (1994) attributed the extreme length of subjects to three "pressures", namely the need to be: (a) precise, (b) concise, and (c) efficient in making claims.

The call by casting rule from the study on aggregation by Dalianis & Hovy (1993) (see section 2.4.6.1) for consistent use of the same nominalized form is not always complied. Summaries are not devoid of still reducible noun phrases. Depending on sentence constructs, and for various reasons, a continuum of expressions from full clause, to complex noun phrase to nominalized form are usually found. Where lexical frequency is used in the selection of sentences, the different linguistic forms sharing the same referent should correctly be gathered together. Consider the example sentence given in Table 5-1. For a frequency that is representative of the concept *larvae*, the cumulative frequency of different forms (i.e. *larvae* and *spider boarder*) of the concept has to be obtained. Document and domain knowledge is required.

5.1.3 Compounding

Compounding or compound word formation involves the adjunction of two or more free morphemes (Crystal, 1997). For our discussion, we consider only nominal compounds. In the most common kind of nominal compound, 'determinative compound', the first element determines the meaning of the second (Bussmann, 1996). Consequently, reordering the elements will alter the meaning of the compound, e.g. *palm oil* vs. *oil palm*. As compounding allows complex ideas to be expressed in a concise way, it is widely used in scientific English (Upjohn *et al.*, 1991:111).

By determining the semantic pattern of compound words, we can formalize the correspondence between a given pattern and its encoded meaning(s), e.g.

Semantic pattern		Encoded meaning	Examples
PLANT PRODUCT	⇔	'product of plant'	*maple syrup, soy sauce*
PLANT OBJECT	⇔	'object made from product of plant'	*rubber shoes*
PRODUCT PLANT	⇔	'plant that produces such a product'	*timber tree, sugar cane*

However, correct reading of elements within multi-word compounds and decoding of its meaning can be tricky. Domain knowledge of more basic compounds is often necessary for correct reading and decoding. Because, **genetic color*, *?abstracting research*, **spider-associated ecology* [* = not acceptable; ? = acceptability is questionable], hence

*(*genetic color*) *variation*	but	*genetic* (*color variation*)
**automatic* (*abstracting research*)	but	(*automatic abstracting*) *research*
*(*spider-associated ecology*) *of insects*	but	*spider-associated* (*ecology of insects*)

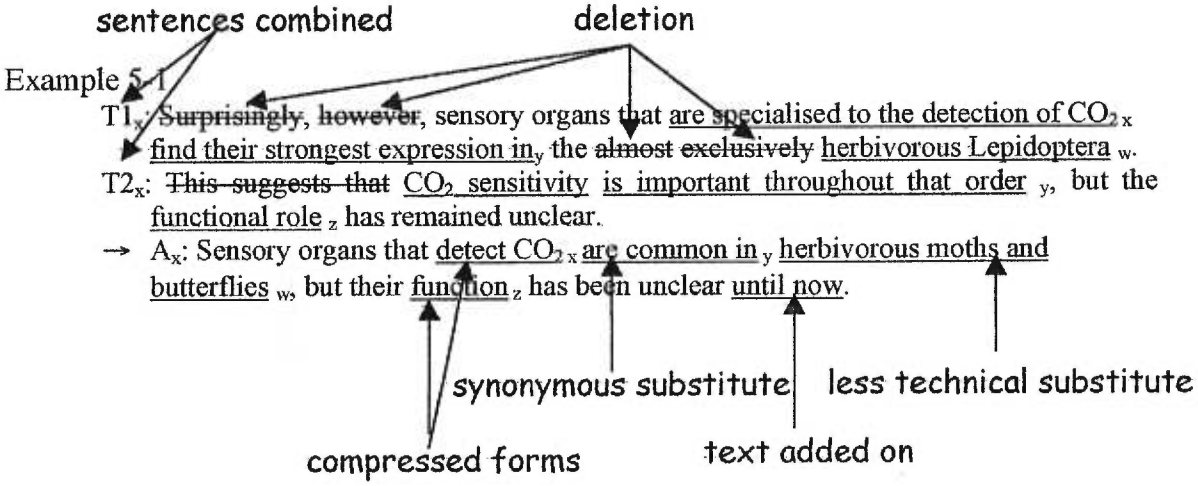
Compound words whose meaning have lexicalized, like terminology, constitute another group of frequently occurring nominal compounds. Such compounds are of little interest to us, and are at best treated as complex lexemes.

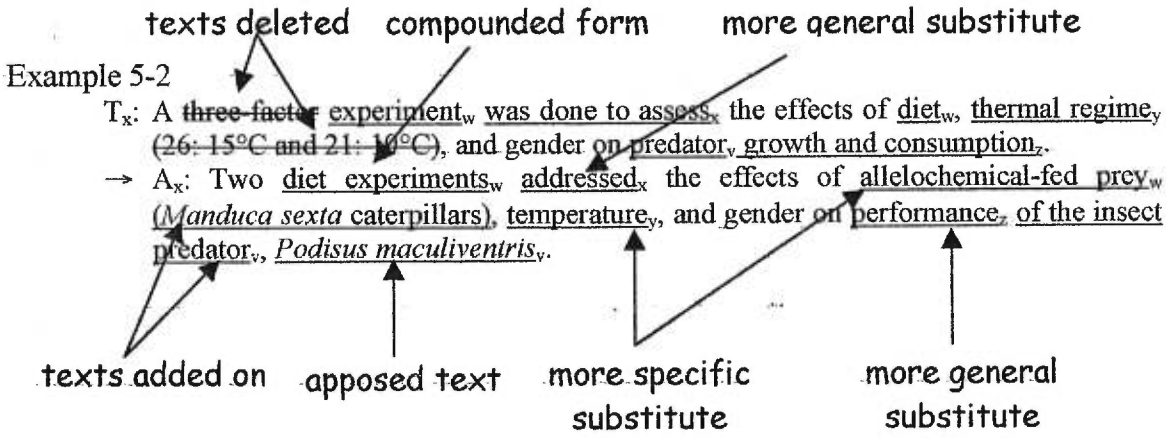
5.2 Condensation Sub-processes in Abstracting by an Author

So far, we have seen some condensation sub-processes proposed or identified by other researchers in summarization or text generation, and some mechanisms the English language has for expressing content concisely. However, what sort of mechanism do authors actually use in abstracting? Are there still other condensation mechanisms in summarization besides those that we have seen?

Without access to how authors condense, we have no recourse but to an indirect deduction by comparing sentences from full text and abstract. Consider the following two examples which were extracted unedited from our study corpus. Subscripts are added to indicate the corresponding units involved.

Full text sentences, $T1_x$ and $T2_x$, were determined to have been the source of information for the sentence A_x in abstract (see section 3.3 identification of ft-sentences). Among the processes identified in abstracting by an author are: (a) deletion, (b) substitution with generic form/expression, which is generalization (c) use of more precise form/expression as a replacement, which is particularization, (d) replacement with an equivalent form, which is substitution with synonym, or synonymous expression, (e) the combining of segments of text, which is aggregation, (f) insertion of additional information, which is the inverse of deletion, (g) apposition of text, which we group with deletion and insertion, but as a group on its own, and (h) use of compressed forms. In sub-section 3.3.3, we saw an example of de-aggregation (see Example 3-6), and also an example of de-nominalization where lexical unit *differences* was verbalized into lexical unit *differ* which is the converse of compression, and which we will refer to as expansion (see Example 3-8).





In Example 5-1 and Example 5-2, the substitutes may be grouped into two types on the basis of form which we have set out in separate columns in Table 5-2. The unit on the left of the arrow is from full text, and that on the right, from abstract.

Table 5-2. Substitution with Synonymous and Compressed Substitutes

Synonymous Substitute	Compressed Substitute
find their strongest expression in → are common in	*functional role → function*
Lepidoptera → moths / butterflies	*~~are specialised to the~~ detection of X → detect X*
was done to assess → addressed	
diet → allelochemical-fed prey	
thermal regime → temperature	
predator growth and consumption → performance	

Consider the first example in the right column. On the basis of redundancy in meaning between adjective *functional* and the noun that it modifies, linguistic unit *functional role* was reduced to just *function*. The operation may be described by a rule such as:

$[A_0(N_x)_{\text{modifier}}\ N_{y\text{-head}}] \rightarrow [N_x]$ | meaning of $[A_0(N_x)\,N_y] \cong$ meaning of $[N_x]$;

In the second example, the linguistic unit *are specialised to the detection of X* is reduced to just its essential units of *detect X*. The operation may be described by the following rule. In both examples, the substitute shares lexeme(s) with the source linguistic unit. We refer to the process as substitution with a compressed substitute, or compression.

BE ADJ $[\ S_0(V)\ [X]\ ]$[45] $\rightarrow$ V [X] | with special constraints on ADJ;

Now, consider the examples in the left column. The operation is not describable by rules. Special lexical resources are required for the operation. Unlike the operation in the right column, none of the lexemes in the source lexical unit in the left column is retained in its substitute.

Because content is compressed into a few indispensable units, as opposed to substitution with synonyms or synonymous expressions where a unit is entirely substituted, we prefer to make a distinction between these two types, and include a fourth category of condensation sub-processes of COMPRESSION. Nominalization is an example of a compression process.

[45] [ ] the brackets indicate embedment. .

5.3 Typology of Condensation Sub-processes and their Definitions

As noted in our own comparative study of abstracting by an author himself in a scientific domain, and also in the work of fellow researchers in Chapter 2, condensation does not implicate strict deletion or re-expression of content in lesser number of words. Units or segments of text may equally be inserted to compact more information into an abstract. A linguistic unit may equally be re-expressed in longer form. We recall our provisional definition of condensation given at the end of Chapter 2 which we give again here.

> CONDENSATION The process of making something that is spoken or written shorter, by not giving as much detail or using fewer words to give the same information, or by augmenting information such that a unit is informatively more compact, or more explicitly expressed.

With the definition as guide, we will set out the categories of condensation sub-processes, and separate out the non-condensation processes. Wherever possible definitions from linguistics will be used, and amended accordingly where necessary. As mentioned above, inverse processes (which run in the opposite direction as condensation processes) are occasionally observed in abstracting. We include their definitions and some examples here for completeness. The aim of this discussion is to gather together operationally similar processes, which we hope will shed light on the units involved for particular condensation processes, and their transformation patterns. Our restrictive study is a start to the latter aim.

A detailed discussion of the linguistic units involved for each condensation process, and their respective inverse process, will be presented later in separate chapters (see chapters 6, 7, 8 and 9) to further elucidate the meaning and complexity of the sub-processes. We discuss below the four categories of condensation sub-processes:

(a) generalization,
(b) deletion,
(c) compression, and
(d) aggregation.

5.3.1 Generalization

In linguistics, generalization commonly refers to change in word meaning over time. However, it has been used in the context of summarization to refer to the replacement of "phrases or clauses with more general descriptions" as proposed by Jing & McKeown (2000:179). By specifying "phrases and clauses" Jing & McKeown (*ibid.*) restrict the definition which may also apply to words. Also, while Jing & McKeown (*ibid.*) specified the substitute to be "more general descriptions", we prefer the substitute to include hypernyms. Generalization was used by Sparck Jones (1999) although she offered no definition. Below is our proposed definition modified from that given Jing & McKeown (2000).

> GENERALIZATION Replacement of a lexical unit or a group of lexical units with its hypernym, or a more general description.
>
> e.g. predatory insects → predatory arthropods
> e.g. predator growth → performance

The examples given are taken from Example 5-1 and Example 5-2 in section 5.2. The relation between a substitute and the unit substituted for may not be evident, and may involve domain knowledge.

The provisional definition of generalization given above covers Maybury's (1995:742) usage of "replacement of a series of events with a single event" which we find acceptable although too restrictive. Generalization is preferred over Maybury's (*ibid.*) proposed term of ABSTRACTION which while appropriate, may be confused with ABSTRACTING in the sense of "writing abstracts".

At the end of Chapter 2, we proposed the use of particularization as the inverse of generalization. The Longman Dictionary of Contemporary English (LDOCE) (1995) defines PARTICULARIZE as "to give the details of something". However, consistent with the definition for generalization, we propose the following definition.

PARTICULARIZATION Replacement of a lexical unit or a group of lexical units with its hyponym or a more specific description.

e.g. Lepidoptera → moths / butterflies

e.g. thermal regime → temperature

In linguistics, the definition for SUBSTITUTION, which is

SUBSTITUTION A term used in linguistics to refer to the process or result of replacing one item by another at a particular place in a structure.

The definition of substitution subsumes the two definitions of generalization and particularization proposed for the context of summarization. We will use this definition from linguistics as the generic definition for generalization and particularization. For substitution, a substitute is not semantically more, or less specific, but semantically equivalent to the substituted linguistic unit.

e.g. find their strongest expression in → are common in

e.g. ten weeks → about three months

In rule form, we write substitution as:

[X] → [Y] | 'X' ≅ 'Y'[46]; X, Y = linguistic units;

and those of generalization and particularization respectively to be

[X] → [Y] | 'X' ⊂ 'Y'; X, Y = linguistic units;

[X] → [Y] | 'X' ⊃ 'Y'; X, Y = linguistic units;

[46] 'X' ≅ 'Y' is to mean that the meaning of X is approximately equal to that of Y. While 'X' ⊂ 'Y' is to mean that the meaning of Y is more general than that of X, 'X' ⊃ 'Y' is to mean that the meaning of Y is less general than that of X.

5.3.2 Deletion

The most obvious way to condense a text is to DELETE linguistic units. To delete is "to remove a letter, word etc from a piece of writing" (LDOCE, 1995). We define deletion as follows.

> DELETION The removal of a linguistic unit Y from another linguistic unit X which is delimited or made more specific by the more dependent unit Y.

e.g. [~~the~~ $_x$ ~~almost~~ $_y$ ~~exclusively~~ $_z$ [herbivorous Lepidoptera]]
→ [ϕ_x ϕ_y ϕ_z herbivorous moths and butterflies],

e.g. [~~however,~~ $_x$ [sensory organs that …]]
→ [ϕ_x [Sensory organs that …]]

e.g. the effects of … thermal [regime$_y$ [~~(26: 15°C and 21: 10°C),~~ $_x$]] …
→ the effects of … [temperature$_y$ ϕ_x], …

Accordingly, INSERTION is defined as the inverse process of deletion.

> INSERTION The adding on a linguistic unit Y to another linguistic unit X, such that the unit X is delimited or made more specific by the more dependent unit Y.

e.g. […experiment] → [[diet] experiments].

e.g. [… has remained unclear]. → [… has been unclear [until now]].

e.g. on [predator growth and consumption$_z$].
→ on [performance$_z$ [of the insect predator]].

APPOSITION describes "a sequence of units which are constituents at the same grammatical level, and which have an identity or similarity of reference" (Crystal, 1997). Reformulating Crystal's definition, we define apposition as follows. We do not categorize apposition with aggregation because no larger unit is formed.

> APPOSITION The positioning of a linguistic unit Y to another linguistic unit X, where X and Y are constituents at the same grammatical level, and have identity or similarity of reference.

e.g. [can inhibit the development of their hosts]
→ [reduce the growth rate of its [host$_i$] [*N. plumipes*$_i$]].

In rule form, we write deletion, insertion and apposition respectively as:

[X[~~Y~~]] → [X ϕ_Y]

[X] → [X [Y]]

and

[X] → [X$_i$] [Y$_i$], | X and Y co-refer as indicated by subscript i;

5.3.3 Compression

The general definition for COMPRESS is "to write or express something using fewer words" (LDOCE, 1995). For our purpose, expressing a linguistic unit "using fewer words" is too vague. Substitution, deletion and aggregation too can involve the use of fewer words. Important in compression is keeping some lexemes from the source unit. Hence, we will qualify the definition given with "lexemes originally found in the source unit" which are **given in bold**, and "essential lexemes" which are **additionally underscored** in the examples given below.

> COMPRESSION The re-expression of a linguistic unit in fewer essential lexemes originally found in the source unit, and in the process. The relation holding between lexemes in the source unit is rendered implicit in the relation holding between the essential lexemes retained.

e.g. **functional role** → **function**

e.g. **foraging by ants** → **ant foraging**

e.g. **none of the 30 distributions** → **no distribution**

We recall example, *ten weeks* → *about three months*, given in sub-section 5.3.1.2 which was considered as a substitution process. Unlike compression, substitution does not retain any of the lexemes originally in the source unit in the substitute. However, as one or more lexemes originally in the source are retained in a nominalized form, nominalization is by our definition a compression process (see example below is taken from Example 3-15 in sub-section 3.4.6).

e.g. **deciding whether or not to lay → the decision to lay additional eggs**

In compression of a unit, the overall meaning is essentially retained. However, the relation between lexemes in the source unit is now implicit in the substitute. We will use the antonym of compression, i.e. EXPANSION, as its converse form.

EXPANSION In expansion, a compressed unit is changed into its full form. The meaning of the unit, or the relation holding between the lexemes constituting if there is more than one, is made explicit.

In rule form, we write compression and expansion respectively as:

[x y z] → [y'] | x, y, y', z = lexical units; y, y' share lexeme Y;
| 'x y z'' ≅ 'y'';

[z] → [x y z'] | x, y, z, z' = lexical units; z, z' share lexeme Z;
| 'z' ≅ 'x y z'';

5.3.4 Aggregation

In sub-section 2.4.3, Dalianis & Hovy (1993) proposed aggregation to refer to the removal of redundancy, and Paice (1981) define it to be the adding of adjacent sentences. To refer to this sub-process in the context of abstracting, the term with the underlying meaning of "combining segments of text" will be retained, but with the stipulations of "removal of redundancy", and "adjacent" taken out. Not only is deletion not necessarily implicated[47], but sentences aggregated need not be adjacent. In the examples below, we indicate aggregation by a plus sign +.

> AGGREGATION The combining of two linguistic units X and Y with the use of an explicit sign, such as a connective (e.g. and or but), a colon or semi-colon, to form a larger unit. Two linguistic units may also be aggregated without the use of such explicit signs.

e.g. [*P. zelicaon* ... is one of the most broadly distributed butterflies in western North America] +
[P. zelicaon is also one of the most polyphagous butterflies, ...]
→ [*Papilio zelicaon* ... is one of the most widely distributed and polyphagous butterflies in western North America]

e.g. [insects] + [spiders] → [arthropods]

e.g. [growth] + [consumption] → [performance]

e.g. [recent study] + [field studies] → [preliminary field observations]

[47] Consider the trivial example where no linguistic unit is deleted: *The elephant is big* + *The mouse is small* → *The elephant is big, but the mouse is small.*

In sub-subsection see 2.4.6.2, Reape & Mellish (1999) proposed two definitions for aggregation, a broad and a narrow definition. The converse of AGGREGATION is DE-AGGREGATION.

DE-AGGREGATION The splitting of a linguistic unit [X ... Y] into two simpler linguistic units X and Y.

In rule form, we write compression and expansion respectively as:

[X] + [Y] → [X ..: Y]

and

[X ... Y] → [X] + [Y]

5.4 Concluding Remarks

In Fig. 5-1 we summarize our proposed typology for condensation sub-processes identified by various researchers, linguistics and our own comparative study for use within the context of summarization. Their inverse processes are also given. Depending on how one chooses to see the process, the typology may accordingly be revised. However, more important to the domain is to identify the range of condensation processes, and to be able to distinguish between the processes, hence, the reason for the their provisional definitions (Aim 2) which are serve as guide to distinguish between them.

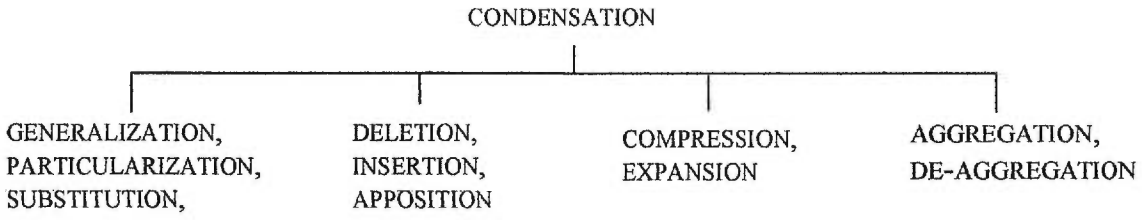


Fig. 5.1. Typology of Condensation Sub-processes

With this we have determined the common linguistic mechanisms applied in abstracting in one specific scientific domain (Aim 1 of our study). However, we still do not know the type or the function of the linguistic units involved for each type, nor the complexity of the problem involved. One condensation sub-process may be accompanied by another sub-process, e.g.

aggregation may be accompanied by generalization. In the next four chapters (6, 7, 8 and 9) we will look into the type and function of linguistic units involved (Aim 3) in each of the four categories of sub-processes starting with SUBSTITUTION. At the same time, the utility of WordNet to summarization will simultaneously be investigated (Aim 4).

Chapter 6

Types of Lexical Substitute in Abstracting

Chapter 6 discusses the first of four groups of condensation sub-processes that we have just seen, GENERALIZATION. In the reformulation of selected content in abstracting, authors may reuse the stems of lexical units of selected sentences[48] (ft-LU) in same or different form, or use in its place, a substitute (ab-LU) that is in some lexico-semantic relation with it: its synonym, hypernym/superordinate, or hyponym, or an expression that is more, or less, precise compared to its ft-LU. The link between lexical units however is not always evident.

To know the type and proportion of substitutes involved, and how the meaning of selected content, ultimately the abstract, is changed in terms of generality, or specificity, as a result of reformulation during abstracting, section 6.1 describes the categorization and quantification of substitution types.

For an indication of the linguistic units involved and the type of knowledge required (Aim 3 of our study), section 6.2 exemplifies and discusses each substitution type in its context. To explore the utility of a ready resource, namely WordNet (WN), in automatic abstracting (Aim 4 of our study), the lexico-semantic relation between lexical units (LUs) involved is simultaneously looked up, and given in the notation of the Explanatory-Combinatorial Dictionary (ECD) in the framework of Meaning-Text Theory (Melcuk *et al.*, 1995; Melcuk, 1996).

Briefly, lexical functions from the ECD are intended to describe lexical relations, and not to encode knowledge. Nevertheless, for our purpose, some WN relations, such as hypernymy and hyponymy which relate concepts, are presented in the lexical function format of the ECD. However, to distinguish WN relations from ECD lexical functions which are in abbreviated forms, e.g. **Syn**(LU), WN relations are given in full, e.g. **Hypernym**(LU).

In the last section, we sum up our observations, operations and the different knowledge resources required for each substitution type. A reason for the difficulty in the implementation of condensation processes lies in the different kinds of knowledge required, specifically domain

and world knowledge, which need to be treated differently. A first step towards alleviating this problem for abstracting is to compile "more stable" knowledge into a special resource for abstracting in a restricted domain. The interplay with other sub-processes, and factors with regards to purpose and readers at whom the abstracts are targeted are other reasons.

6.1 Categorization and Quantification of Substitution Types

To categorize and quantify the types of substitution, a comparison of the form and meaning of ab-LUs and their corresponding ft-LUs was carried out. The categorization separates out the freely inter-substitutable units from those that are not. By "freely inter-substitutable units" we mean those units whose lexical information is based only on linguistic knowledge of the language or sublanguage, and is likely to be coded in a lexical resource. Non-freely inter-substitutable units involve those substitutions which require various considerations, such as context, and experimental and/or world knowledge.

For this particular study, only a random sample of 55 ab-sentences (10% of corpus) from eleven documents were analyzed. The results were not significantly different even with a smaller corpus. We describe first the four categories of substitutions (Type I to Type IV), and then report the quantification.

6.1.1 Categorization

For our purpose, a LU can be an affix, e.g. *–free*, or a group of words, e.g. *are specialised to the detection.* In the latter complex LU, the lexeme of the most important word in the unit, i.e. DETECT, is taken to be its lexeme. A substitution is categorized based on identity in lexeme and inter-substitutability of the LUs. The guidelines for categorization are described as follows.

6.1.1.1 Type I: Identical, Inflected or Derived Forms

Consider a ft-LU such as *examined.* If its replacement is *examined* or *examines*, then the ab-LU is its identical or inflected form respectively, and *examination*, its derived form. A substitute is a Type I, if its stem is unchanged. As far as the LUs are concerned, a Type I substitution may be

[48] Which are ft-sentences identified to have been used in writing abstracts (see section 3.3).

carried out independent of context, i.e. the LUs are freely inter-substitutable. The relation between the ft- and ab-LUs is strictly linguistic.

6.1.1.2 Type II: Synonyms

Although of different lexemes, an ab-LU and its ft-LU are sufficiently inter-substitutable independent of context to be considered as synonyms (**Syn**).

e.g. *class* (ft-LU) → *category* (ab-LU) **Syn**(class) = category (WN);
leaf → *foliar*

Linguistic and domain-related knowledge are required for Type II substitutions. While general synonyms may be found in a thesaural resource such as WN, domain-related synonyms need to be compiled from empirical corpus studies. A substitute and its ft-LU need not be of the same part of speech to be considered as synonyms, and it is not necessary for a relation between LUs to be found in WN for a substitution to be considered to be of a particular type.

6.1.1.3 Type III: Document Synonyms and Approximate Equivalents

Unlike the previous two types, Type III substitutions are not freely inter-substitutable, but are restricted to the context of a document (i.e. the substitutes are document synonyms), except for numerical expressions which are generalizable to some extent. To effect Type III substitutions, experimental and non-technical knowledge are often required. Type III substitutes may be divided into three groups:

(a) technical words: e.g. *kleptobiont* → *kleptobiotic spider*; *Lepidoptera* → *moth*;
(b) non-technical words: e.g. *vulture* → *bird*; and
(c) numerical expressions: e.g. *10 weeks* → *about 3 months*.

6.1.1.4 Type IV: Complicated Substitutes

The lexical relation between a Type IV substitute and its corresponding ft-LU is not evident. Consider ab-LU *host* and its ft-LU *fruit*. The LUs are not lexically related. It is only from experimental knowledge that ft-LU *fruit* is known to be host to weevils that attack the fruit. Consider complex ft-LU *was significantly lower than* and ab-LU *was a significant departure*

from. Their respective lexemes LOWER and DEPARTURE are weakly related. To arrive at an ab-LU, various kinds of knowledge and manipulations may be involved.

6.1.2 Quantification

The preceding one-way categorization gives an indication of the knowledge needed for each substitution type, but no hint about how a LU has changed in terms of semantic precision as a result of substitution. And so, on the basis of divergence in meaning of an ab-LU from that of its ft-LU, each substitution is further classified as: (a) unchanged, (b) more general or (c) more specific. We will now illustrate this two-way categorization using ft-ab sentence matches.

6.1.2.1 Sub-categorizing a Substitution Type

Example 6-1 is a ft-ab sentence match from our study. Corresponding ft- and ab-LUs have been given the same subscript. Using a working table such as Table 6-1, ft- and ab-LUs were placed in their appropriate columns, and their substitution type and change in meaning determined.

Example 6-1

Full text (oec2- 97110539)	Abstract
T1a: *Surprisingly, however, sensory organs*$_{p}$ *that are specialised to the detection*$_{q}$ *of* CO_{2r} *find their strongest expression in*$_{s}$ *the almost exclusively*$_{t}$ *herbivorous*$_{u}$ *Lepidoptera*$_{v}$. [I-1-7]	A1a: *Sensory organs*$_{p}$ *that detect*$_{q}$ CO_{2r} *are*$_{s}$ *common in*$_{t}$ *herbivorous*$_{u}$ *moths*$_{v1}$ *and butterflies*$_{v2}$,
T1b:*This suggests that* CO_2 *sensitivity is important throughout that order, but the functional role*$_{w}$ *has remained*$_{x}$ *unclear*$_{y}$. [I-1-8]	A1b: *but their function*$_{w}$ *has been*$_{x}$ *unclear*$_{y}$ *until now.* [A-1-1]

Ab-LU *Sensory organs*$_{p}$ is identical to its ft-LU *sensory organs*$_{p}$. The substitution is a Type I. Because the stems of Type I substitutes are unchanged from their respective ft-LUs, their meanings are considered to be unchanged.

Table 6-1. Working table for categorization of substitution types for Example 6-1

ft-LU	ab-LU	Substitution type	Change in meaning		
			unchanged	more general	more specific
sensory organs$_{p}$	*Sensory organs*$_{p}$	I	+		
are specialised to the detection$_{q}$	*detect*$_{q}$	I	+		
CO_{2r}	CO_{2r}	I	+		
find their strongest expression in$_{s}$	*are*$_{s}$	IV		+	
the almost exclusively$_{t}$	*common in*$_{t}$	IV		+	
herbivorous$_{u}$	*herbivorous*$_{u}$	I	+		
Lepidoptera$_{v}$	*moths*$_{v1}$	III			+
Lepidoptera$_{v}$	*butterflies*$_{v2}$	III			+
functional role$_{w}$	*function*$_{w}$	II	+		
has remained$_{x}$	*has been*$_{x}$	IV		+	
unclear$_{y}$	*unclear*$_{y}$	I	+		

+ = meaning of ab-LU with respect to ft-LU is unchanged/ more general / more specific.

Ab-LU *detect*$_{q}$ shares the lexeme DETECT of its ft-LU *are specialised to the detection*$_{q}$. As the stem in ft-LU is retained, the meaning in ab-LU is taken to be unchanged. The operation in question is more precisely substitution with a compressed form. As we are interested in the semantic change of lexical units during abstracting, we will for the present ignore the compression process operating. COMPRESSION will be discussed in Chapter 8.

Ft-LU *functional role*$_{w}$ and its corresponding ab-LU *function*$_{w}$ are inter-substitutable out of context. The substitution is a Type II because *role* and *function* do not share the same stem: **Syn**(role) = function (WN). As with ft-LU *are specialised to the detection*$_{q}$ in the previous paragraph, one may arguably say that *function*$_{w}$ is the result of compressing *functional role*$_{w}$, or the result of deletion of redundant unit *functional*, followed by substitution of *role* with its synonym *function*. Like Type I, Type II substitutes are considered to be unchanged in meaning from that of its ft-LU.

Now, consider ab-LUs *moths*$_{v1}$ and *butterflies*$_{v2}$ and their common ft-LU *Lepidoptera*$_{v}$. The ab- and ft-LUs are not freely inter-substitutable. A hypernym may not substitute for its hyponym, except in restricted contexts. While not all lepidopterans are moths, all moths and all butterflies are lepidopterans. *Moth* and *butterfly* are CO-HYPONYMS to *lepidopteran*: **Hyponym**(lepidopteran) = moth, butterfly. The substitution is a Type III, and the change in meaning is to one that is more specific with respect to that of its ft-LU.

The link between complex ft-LUs *find their strongest expression in*$_{s}$ and *the almost exclusively*$_{t}$, and their corresponding ab-LUs, *are*$_{s}$ and *common in*$_{t}$ is not evident. The substitutions are classed as Type IV. Both ab-LUs are less specific in meaning than their ft-LUs. Hence, the change in meaning classed as "more general" for both (see Table 6-1).

Now, let us consider a few more ft-LUs from another ft-ab sentence match, Example 6-2 (given earlier as Example 3-1 in sub-section 3.3.1).

Example 6-2

Full text (bes2-9638145)	Abstract
T7: *When the number of sexual offspring*$_{w}$ *was adjusted for colony size, slave-added colonies*$_{x}$ *had*$_{y}$ *significantly more sexuals*$_{z}$ *than the controls* (*one-tailed test, as the only expected change after adding food or slaves is an increase of sexuals*). [R-2-3]	A7: *When colony size was adjusted to the number of sexual offspring*$_{w}$, *the treatment colonies*$_{x}$ *produced*$_{y}$ *significantly more sexual offspring*$_{z}$ *than the controls*. [A-1-7]

Consider the following two pairs of LUs: ab-LU *treatment colonies*$_{x}$ and its corresponding ft-LU *slave-added colonies*$_{x}$, and ab-LU *sexual offspring*$_{z}$ with ft-LU *sexuals*$_{z}$. For both pairs, the corresponding LUs are inter-substitutable, but only within the restricted context of the document. Both ab-LUs are Type III substitutes. However, while the meaning of ab-LU *treatment colonies*$_{x}$ is more general than the meaning of ft-LU *slave-added colonies*$_{x}$, the meaning of ab-LU *sexual offspring*$_{z}$ is more specific than that of its ft-LU *sexuals*$_{z}$ (see Table 6-2).

Consider the ab-LU *produced*$_y$ and its corresponding ft-LU *had*$_y$. The link between the two LUs is not evident. The substitution is classed as Type IV. Ab-LU is more specific in its meaning than ft-LU.

Table 6-2. Working table for categorization of substitution types for Example 6-2

ft-LU	ab-LU	Substitution type	Change in meaning		
			unchanged	more general	more specific
colony size	*colony size*	I	+		
adjusted	*adjusted*	I	+		
sexual offspring$_w$	*sexual offspring*$_w$	I	+		
slave-added colonies$_x$	*treatment colonies*$_x$	III		+	
had$_y$	*produced*$_y$	IV			+
significantly more	*significantly more*	I	+		
sexuals$_z$	*sexual offspring*$_z$	III			+
controls	*controls*	I	+		

+ = meaning of ab-LU with respect to ft-LU is unchanged/ more general / more specific.

6.1.2.2 Distribution of Substitution Types

From this two-way categorization of substitution, we obtained a rough quantification of the type and proportion of substitutes in abstracting, which is summarized in Table 6-3.

Table 6-3. Distribution$^+$ of substitution types

Type	Meaning of ab-LU with respect to ft-LU			Total
	unchanged	more general	more specific	
I	195	-	-	195 (54%)
II	17	-	-	17 (5%)
III	3	39	38	80 (22%)
IV	-	41	31	72 (20%)
Total	215 (59%)	80 (22%)	69 (19%)	364

$^+$ = Data based on a sample of eleven documents randomly selected from corpus;
ft-LU = lexical unit in full text; ab-LU = lexical unit in abstract;

The greatest proportion of substitutions are with identical, inflected or derived forms (Type I), while one-fifth of substitutions are with quasi-synonymous forms and equivalents (Type III). Overall, meaning is re-expressed in slightly more general than specific terms during abstracting. We will now look at the linguistic unit involved in the various substitution types, and the kinds of knowledge required.

6.2 Substitution Types

Condensation is a complex process, more than one ft-sentence may be the source of information for an ab-sentence (see sub-section 3.3.2), and more than one condensation sub-processes may operate on a LU. After teasing apart the sub-processes, we focus on substitution and ignore all other processes that might also be operating.

Extracted examples are given unedited in the following format. However, texts that are marginal to the discussion are excluded, and their omission indicated by three consecutive dots (...). The segments of text in discussion are underscored.

(0) <ft-sentence> [location code of ft-sentence]

→ <ab-sentence> [location code of ab-sentence; jjjj-yy_vv_ppp]

LEXICO-SEMANTIC_RELATION(ft-LU) = ab-LU

For many researchers in automatic summarization, the interest is if transformations observed in condensation can be operationalized. Consider *dramatic differences* → *differ dramatically*, and *dramatic differences* → *differ dramatically*, which may be formalized by a rule (see below) consisting of a mix of linguistic units given in abbreviated form, e.g. N(oun), Adv(erb), etc., some of which are in ECD notations, e.g. S_0(V), Adv_0(A), etc. [S_0(V) = substantive form derived from verb, V; Adv_0(A) = adverb derived from adjective, A].

$[A_{modifier}\ S_0(V)_{head}]_{NP} \rightarrow [V_{head}\ Adv_0(A)_{modifier}]_{VP}$

The subscripts indicate the syntactic functions of the linguistic units. The two syntactic functions used are HEAD and MODIFIER (abbreviated modf).

6.2.1 Type I

About 55% of ab-LUs were identical, inflected or derived forms of their ft-LUs (see Table 6-3). Type I substitutions are convenient ways of expressing content differently, but accurately. The substitution involves linguistic knowledge, and the LUs involved are from major parts of speech.

6.2.1.1 Identical and Inflected Forms

In Type I substitutions with inflected forms, a segment of text may be retained almost in its entirety except for minor changes in form and ordering. There is no obvious reduction in terms of words.

(6-1) a conspicuous tuft of bristles and dark pigmentation on the tibia and patella of the first pair of legs [I-5-3]

→ dark pigmentation and tufts of bristles on the tibiae of their forelegs.

[A-1-1; bes1-9638017]

(6-2) In the arena allowing the transmission of both vibratory and visual signals, ... [R-2-1]

→ in arenas that allowed both visual and vibratory signal transmission ...

[A-1-5; bes1-9638017]

6.2.1.2 Derived Forms

As in substitutions with inflected forms, there is no obvious reduction in terms of words in Type I substitutions with derived forms.

(6-3) Motivation-based models ... currently remain the only theoretical framework for ... [I-1-6]

→ Motivation-based models dominate current theory concerning ...

[A-1-1; oec1-99120274]

In cases where an adverb is replaced by its prepositional adverbial form, there is an insignificant increase in number of words. Not all adverbs may be paraphrased as prepositional adverbials. An ECD once available should provide this information.

(6-4) Specialized arenas were used to isolate experimentally$_{Adv}$ ... [M-4-1]

→ In behavioral experiments$_{N}$ that paired ... in arenas ... [A-1-5; bes1-9638017]

Unlike Type I substitutions with inflected forms, the regular co-occurrence of certain types of derivations, e.g. verbalization with adverbalization (see example (6-5)), and nominalization with adjectivalization give convenient patterns for reformulating content (see example (6-6)).

(6-5) Our results show that prey may exhibit dramatic$_A$ differences$_N$ in quality. [D-1-2]

→ The results … suggest that prey species … may differ$_V$ dramatically$_{Adv}$ in their suitability as food for … [A-1-12; oec2-99119191]

$[A_{modf}\ S_0(V)_{head}]_{NP} \rightarrow [V_{head}\ Adv_0(A)_{modf}]_{VP}$ | with undetermined constraints on A and V;

(6-6) Many species … can potentially$_{Adv}$ compete$_V$ with and prey upon each other. [I-1-2]

→ Spiders and ants are potential$_A$ competitors$_N$ and mutual predators. [A-1-1; oec2-97109313]

(6-7) aphid foundresses do not discriminate$_V$ actively$_{Adv}$ between kin and non-kin [D-1-1]

→ active$_{Adj}$ kin discrimination$_N$ by *T. coweni* foundresses [A-1-5; bes1-9843095]

$[Adv_0(A)_{modf}\ V_{head}]_{VP} \rightarrow [A_{modf}\ S_0(V)_{head}]_{NP}$ | with undetermined constraints on A and V;

With nominalization, a different semantic category is emphasized. For example, focusing on the agent, e.g. *competitors*, instead of the act, e.g. *compete* in example (6-6). The respective syntactic functions of the linguistic units remain unchanged after substitution, i.e. the unit that was head of a VP remains a head, but of a NP, and the unit that was modifier to a verb, remains a modifier, but to a noun. Such patterns are of interest to researchers in automatic summarization, and the ECD lexical functions are convenient means for their description. Nominalization with compounding can result in a significant reduction in text (see example (6-7)).

A second pattern for reformulating selected content allows focus to be shifted onto a different stem, consequentially, changing their syntactic functions with respect to one another. The lexeme functioning as head is now a modifier, and vice versa.

(6-8) sexual trait size … should be largest in the most viable$_A$ males$_N$ [I-1-1]

→ the relationship between male sexual trait size and male$_N$ viability$_N$ [A-1-1; bes1-9946123]

$[A_{modf}\ N_{head}]_{NP} \rightarrow [A_0(N)_{modf}\ S_0(A)_{head}]_{NP}$ | with undetermined constraints on A and N;

(6-9) biased$_A$ sex ratios$_N$ are the result of an overproduction of female embryos. [I-2-4]

→ sex ratio$_N$ bias$_N$ of the African social spider *Stegodyphus dumicola* Pocock[49] is the result of an overproduction of female embryos. [A-1-1; bes1-9946237]

$[A_{modf}\ N_{head}]_{NP} \rightarrow [\ N_{modf}\ S_0(A)_{head}]_{NP}$ | with undetermined constraints on A and N;

While we did not make an actual count for both patterns in the corpus, we have about a dozen of examples.

6.2.2 Type II

Type II substitution which replaces ft-LUs with synonyms constitutes only about 5% of substitutes. Examples from both major and minor parts of speech were found.

6.2.2.1 Major parts of speech

For Type II substitutions involving major parts of speech, fewest examples were found for nouns, and most for adjectives and adverbs (see examples (6-10) to (6-15); for more examples, see Appendix V Table A5-1). The synonyms for most ft-LUs were found in WN. With verbs, the relation more often than not involves a troponym (see example (6-12)).

(6-10) a first class of fruits [D-1-1]

→ the first category of hosts [A-1-9; oec2-98114382]

Syn(class) = category (WN)

(6-11) By retreating underground, Argentine ants avoided ... parasitoid attack. [R-6-3]

→ Argentine ants ... returned underground ... [A-1-6; oec2-98117420]

Syn(retreat) = return (WN)

---

[49] A scientific name consists of <name of genus> <name of species>(<name of author of scientific name>). The genus and species which are in given in italics or underscored, may appear in the following format of *Stegodyphus dumicola*, *S. dumicola*, or *Stegodyphus* sp.. The name of author of a scientific name may optionally be given within or without parentheses, e.g. *Stegodyphus dumicola* Pocock or *Stegodyphus dumicola* (Pocock).

(6-12) by … cutting … with dissecting scissors. [M-7-3]
→ by manually removing; [A-1-7; oec1-97111570]

Troponym (remove) = cut off (WN)

(6-13) same clinal trend [M-7-3]
→ similar clinal trend [A-1-4; bes1-9946025]

Syn(same) = similar (WN)

(6-14) can potentially compete with and prey upon each other. [I-1-2]
→ are potential competitors and mutual predators. [A-1-1; oec2-97109313]

mutual = towards each other (LDOCE)

(6-15) These results suggest that undertakers and guards are somewhat developmentally advanced … [R-7-5]
→ These results suggest that undertakers and guards may be slightly developmentally advanced … [A-1-8; bes2-9741151]

Syn(somewhat) = slightly (WN)

While the synonyms of most ft-LUs were found in WN, a special lexical resource (SR) to link technical lexical units from restricted domains to non-technical words, such as in example (6-16) needs to be compiled empirically to complement a general thesaurus (see Table A5-2a in Appendix V for more examples on Type II substitution).

(6-16) altitudinal gradient in leaf nitrogen [D-6-6]
→ altitudinal gradient in foliar chemistry [A-1-8; oec2-98117133]

Syn(leaf) = foliage (WN)
$\mathbf{A_0}$(leaf) = foliar (SR)

6.2.2.2 Minor parts of speech

Like major parts of speech, minor parts of speech were also substituted with semantically equivalent units during abstracting. The synonyms of most minor parts of speech in our study were found in WN which in general does not encode function words, although one might be able to find some adverbial phrases, e.g. *on the contrary*.

(6-17) wounds ... could exacerbate the rate of desiccation, ... [I-2-3]

→ Wounds ... might exacerbate this problem. [A-1-2; oec1-98115184]

Syn(could) = might (WN)[50]

(6-18) In contrast, over the 3-year study a total of 84 male *P. phalaenoides* were captured on baited traps, while only 2 were captured on control traps. [R-1-2]

→ Each year during our 3-year study, significantly more *P. phalaenoides* were captured on sticky traps ... than on unbaited control traps. [A-1-9; oec1-97112572]

over ≅ during (SR) | LU precede TEMPORAL nouns;

(6-19) both predators [D-4-4]

→ two predators [A-1-19; oec2-97117258]

both = two considered together[51] (WN)

both ≅ two (SR)

(6-20) During our observations, none of the many other ant species interacting with Argentine ants was attacked by *P. pusillum*. [D-3-5]

→ *Pseudacteon* parasitoids commonly attacked Argentine ants, but not other ant species, [A-1-5; oec2-98117420]

none = not at all (WN)

[50] WordNet considered *could* and *might* as verbs.
[51] Taken from the accompanying definition.

(6-21) unlike fluctuating asymmetry, … asymmetry in leg tufts of the wolf spider *Schizocosa ocreata* arises from … [I-1-4]

→ In contrast to fluctuating asymmetry resulting from developmental instability, leg tuft asymmetry in *S. ocreata* [A-1-7; bes1-9638253]

unlike ≅ ⌈in contrast to⌉ (SR) ⌈phrase⌉ to be treated as a single unit;

(6-22) bristles … on the tibia … of the first pair of legs of mature male *S. ocreata*, … [I-5-3]

→ bristles on the tibiae of their forelegs. [A-1-7; bes1-9638017]

fore- ≅ ⌈first pair⌉ (SR)

(6-23) a … higher abundance in ant-free trees [R-3-5]

→ increased … by 1.5- to 1.8-fold in trees without ants [A-1-6; oec2-97109313]

N-free ≅ without N (SR)

If abstracting systems are to carry out Type II substitutions on minor parts of speech and some common expressions as authors do, then a compilation of synonymous function words and expressions, and the contextual constraints on their application (see example (6-18)), should be accessible as another supplementary resource during abstracting, especially if the LUs involve "abstractese" forms (see examples (6-21) and (6-22)). The LUs encoded in the resource may include affixes and their equivalent forms (see examples (6-22) and (6-23)) for restricted domains. In sub-subsection 6.2.2.1 we mentioned a special lexical resource to linked technical words to non-technical words. In (6-21), we note that the unit replaced is shorter than the unit replacing it.

6.2.3 Type III

Unlike the previous two types, Type III substitutes may not properly be considered as synonyms. The substitution is unidirectional, and restricted to the document in question. To effect the substitution, various kinds of knowledge is required. We discuss the substitutes under three groups: (a) technical terms, (b) general words, and (c) numerical expressions. About one-fifth of substitutions belong to Type III in our corpus sample.

6.2.3.1 Technical Terms

Authors were observed to use fewer technical words as substitutes during abstracting. While the relation between a general word, e.g. *experiment*, and a technical word, e.g. *control* or *treatment*, may be clear to a domain reader, the words themselves may not be related lexically, or as they should be, in a general thesaurus. A special resource describing possible relations between general and technical words is necessary to supplement a general thesaurus in abstracting documents from the domain. As there are just so many general words in natural language that can be used as substitutes for technical terms, the domain investigated has its own sublanguage device for creating document synonyms. We divide substitutes for technical words into two groups: (a) domain-related substitutes and (b) document synonyms.

(a) Domain-Related Substitutes

Authors showed a predilection for general words as substitutes for technical words. While some substitutes may not be found in WN (see example (6-24)), for others found, there remains, as for general words, the problem of which unit to use, and at which level of hypernymy.

(6-24) on inducing ant recruitment, we monitored ant numbers on control and treatment *C. obtusifolia* leaves ... [M-4-1]

→ in eliciting an induced response in two experiments. [A-1-9; bes2-9945047]

Hypernym(control) = experiment (SR)

Hypernym(treatment) = experiment (SR)

Consider for example (6-25), the two possible synsets of lexical units, {shed, cast, ...} and {remove, take, take away} from different levels of generality for the word *autotomize*. Domain knowledge required to know which form is the more appropriate to use: *shedding*? *removal*? In the actual replacement, *loss* which is the result of *autotomize*, more than domain knowledge to carry out.

(6-25) a consequence of prior foreleg autotomy and regeneration ... [I-2-3]

→ a consequence of leg loss and regeneration ... [A-1-3; bes1-9945087]

Hypernym(autotomize) = {shed, cast, ...}, {remove, take, take away} (WN)

$\mathbf{S_{\theta}}$ (autotomize) = autotomy (SR)

$\mathbf{S_{res}}$(autotomize) = loss (SR)

In example (6-26), the relation is even less predictable. While WN may give *offspring* as a coordinate term for *embryo* (i.e. LUs sharing the same hypernym), domain and/or experimental knowledge together with other considerations are required to decide which among the possible coordinate LUs to use.

(6-26) sex of individual embryos [I-4-4]

→ sex of individual offspring [A-1-3; bes1-9946237]

Coordinate(embryo) = offspring (WN)

Technical words are not always substituted with general words. Sometimes general words are substituted with technical terms. In example (6-27), generalization is accompanied by aggregation. When LUs are simultaneously aggregated, a reduction in number of words is implicated (see example (6-27)). However, in other examples (see examples (6-28) and (6-29)) there is no apparent reduction in number of words.

(6-27) The abundance of predatory insects and spiders was at least four times greater in … [D-3-1]

→ Predatory arthropods were 4 times more abundant in … [A-1-8; oec2-97112081]

Hypernym(insect) = arthropod (WN);
Hypernym(spider) = arthropod (WN)

(6-28) arises from a single event – regeneration of a leg lost during development – most likely from aggressive encounters with other spiders or predators. [I-1-4]

→ leg tuft asymmetry in *S. ocreata* most likely arises from a single event during ontogeny – … [A-1-7; bes1-9638253]

Syn(development) = ontogeny (WN)

(6-29) from its sibling species *Schizocosa rovneri.* [I-3-3]

→ its sibling congener[52], *Schizocosa rovneri,* [A-1-7; bes1-9638017]

Coordinate(species) = congener (¬WN; SR)

[52] Belonging to the same genus or family.

In example (6-30), a mix of domain and experimental knowledge is required to fill in the text omitted in ft-LU *feed on branches* which should read as *feed on the leaves of branches*, before substitution with *folivore*. The implicit "part" replaces the "whole". Besides "restoring" the implicit text, there is also a shift in focus from the act of feeding to the agent.

(6-30) tussock caterpillars forced to feed on branches that had been damaged earlier in the season by *P. virginalis* grew less rapidly … [I-5-3]

→ feeding caused by either of these two folivores could reduce the performance of the other species. [A-1-2; oec1-99120268]

Hypernym(caterpillar) = folivore (¬WN; SR)

(b) Document Synonyms

During abstracting, technical terms were commonly substituted by special document synonyms. The meaning implicit within a technical ft-LU is extracted and made explicit in a generic head, and in the relation holding between a modifier derived from the ft-LU and the generic head. We refer to this process as EXPANSION to oppose it to COMPRESSION (see sub-section 5.3.3).

Consider technical term *kleptobiont* in example (6-31). When reformulated as *kleptobiotic spider*, the meaning implicit in the technical ft-LU is now made explicit, and is more susceptible to comprehension because of the non-technical generic head. While easily recognized and accepted as synonyms within the document, these "expanded" two-word substitutes may not be used freely as domain synonyms: not all kleptiobionts are spiders.

(6-31) their kleptobionts *Argyrodes antipodianus* Cambridge [I-4-3]

→ a small kleptobiotic spider [A-1-1; oec1-97111570]

In example (6-32), *cannibalism* can equally be expanded into *cannibalistic activity* or *cannibalistic act* by post-posing it with *activity* or *act*. Depending on the hypernym used, the emphasis is different.

(6-32) we could not conclude that … cannibalism would be different in various size/age classes of *P. agrestis* juveniles. [R-2-7]

→ but absolute size/age of an individual could not predict the occurrence of a cannibalistic event. [A-1-8; bes1-9945349]

Hypernym(cannibalism) = activity (WN)

This expansion of technical terms in this way may be expressed in formulaic form as follows. However, the problem that remains is which generic to use as head.

e.g. X → [X'$_{modf}$ Y$_{head}$] | X, X', Y are lexical units; X' as a derived form of X;

In our study corpus, there were more examples of such SEMANTIC EXPANSIONS, where the implicit meaning is extracted and made explicit, than SEMANTIC COMPRESSIONS, where the meaning is packed into a single lexical unit (see examples (6-33) and (6-34)). While examples (6-33) and (6-34) may be considered as substitution with compressed forms with redundancy removed, no redundancy was removed in examples (6-35) and (6-36).

(6-33) Motivation-based models ... currently remain the only theoretical framework for ... [I-1-6]

→ Motivation-based models dominate current theory concerning ... [A-1-1; oec1-99120274]

Hypernym(framework) = theory (WN)

(6-34) but the functional role has remained unclear. [I-1-8]

→ but their function has been unclear until now. [A-1-1; oec2- 97110539]

Syn(role) = function (WN)

(6-35) to assess ... dietary composition [I-1-4]

→ to gain insight into its ... diet; [A-1-3; oec1-99120304]

(6-36) the impact of exotic enemies on populations of native insect species [I-1-4]

→ the extent to which exotics have acquired native hosts; [A-1-2; oec1-97112566]

Within the full texts studied, there were many noun phrases sharing this [MODIFIER + GENERIC]$_{NP}$ construct (see examples (6-37) and (6-38)), except that the modifier is a nominalized event noun (such as *cutting*, *mating*), or a name of entity: common name (*psychodid*, *chrysomelid*) or scientific name (*Azteca*, *Cecropia*). In the presence of a generic head, the meaning of EVENT nouns which are grammatically ambiguous (between noun and verb), and the referent of technical and scientific names which are not explicit from their wordforms, are made

explicit. The meaning of linguistic unit *cutting behavior* is easier to understand compared to *cutting*, and likewise with *psychodid fly* compared to *psychodid*. Even though redundant, because of overlapping semantics (event nouns and their generic head share the same semantics), or because of co-referentiality (a common/scientific name and its generic share the same referent), such generic heads are rarely deleted during abstracting, except in the case where it is a scientific name. Either may be deleted.

Event nouns, if not already with a generic head, were post-posed with a generic word during abstracting thereby explicating its meaning (see examples (6-37) and (6-38)).

(6-37) males modify their vocal behavior in different ways or to different degrees depending on the frequency of an opponent's call. [D-2-7]

→ the frequency of an opponent's calls elicits a differential modification of calling behavior, [A-1-7; bes2-9945177]

Hypernym(call) = activity (WN)

(6-38) "area-restricted search": after foraging on a highly rewarding food source, foragers tends to move a small distance. [I-1-2]

→ The naive bees exhibited area-restricted search behavior, ... [A-1-7; bes2-9639381]

Hypernym(search) = activity (WN)

The choice of word to use as generic head is tricky not only because experimental knowledge is often required, but also because of the multiple hypernyms to choose from. However, in the corpus investigated, hypernyms commonly used with event nouns are *behavior* and *activity*: *cutting behavior*, *mating activity*. For common names or scientific names, the generic is determined by its referent. For example, the generic for *psychodid* is fly, and for *chrysomelid*, beetle. The generic for *Azteca* is ant, and for *Cecropia*, tree.

While the information in WN is not ready-made for abstracting, the necessary information is indirectly available in the definitions, or via other intermediate hypernyms.

Hypernym(chrysomelid) = beetle (WN);
psychodid = a fly of the family Psychodidae (WN);
Cecropia = large genus of tropical American trees (WN)

6.2.3.2 General Words

In the previous sub-subsection, we discussed the types of Type III substitutes for technical words. We will now discuss Type III substitutes for general words.

In abstracting, only the gist of document content is reported. To this end, hypernyms and holonyms are clearly preferred. We discuss them below.

(a) Hypernyms

In sub-section 6.2.2, Type II substitutions with synonyms for general words rarely involve nouns. Here in Type III substitution, the category of linguistic units more commonly substituted with their hypernyms are nouns. To effect substitutions with hypernyms, which is GENERALIZATION, linguistic and domain/world knowledge are often required. However, note the fuzzy boundary between what is considered as linguistic knowledge and what is considered as world knowledge. In examples (6-39) and (6-40), the knowledge involved is linguistic, and in example (6-41), the knowledge involved is considered as world knowledge.

(6-39) *L. sclopetarius* constructed their webs on … of the footbridge. [M-1-2]
→ *Larinioides sclopetarius* … frequently builds webs on bridges. [A-1-1; bes1-9946043]

Hypernym(footbridge) = bridge (WN)
Hyponym(bridge) = drawbridge, viaduct … (WN)

(6-40) the remaining 22 were in areas of light or no snowfall. [M-2-2]
→ in *S. miscanthi* in regions of light or no snowfall.. [A-1-3; bes1-9946025]

Hypernym(area) = region (WN)

(6-41) Large numbers (15+) of vultures ... also began soaring shortly thereafter, ... [R-3-7]

→ the altitude of bird flight began increasing rapidly, ... [A-1-6; oec2-99118381]

Hypernym(vulture) = bird (WN)

Hyponym(bird) = flightless bird, parrot, aquatic bird, ... (WN)

As non-pertinent co-hyponyms are included in generalization, imprecision is inadvertently introduced during abstracting. In example (6-42), the hypernym *bird*, which refers to *vultures* and other aerial birds referred to in the document, after generalization semantically includes other flightless and aquatic birds that are not part of the study. Although selected content is reformulated with less precision, comprehension is not significantly affected when Type III substitutions are restricted to the context of a document.

Besides the use of lexical units in generalization, we illustrate below some grammatical means of making general statements. In example (6-42), definite determiner *the* is substituted with indefinite determiner *a*, and in example (6-43), indefinite determiner *a* is substituted with indefinite quantifier *any*.

(6-42) the negative binomial may be generated either by true contagion ... [I-2-3]

→ A negative binomial distribution may be generated by either true or false contagion. [A-1-10; oec2-98114382]

(6-43) a parasitized host has some chance of survival. [D-1-4]

→ the probability of superparasite survival in any superparasitized host. [A-1-2; bes1-9639061]

To generalize, a determiner or quantifier may equally be omitted (see example (6-44)).

(6-44) by conducting an experiment in which symmetric males were paired with females before and after experimental removal of one of the tufts by shaving. [M-5-1]

→ Experimental removal of $\phi_{one\_of\_the}$ tufts from one leg of previously successful symmetric males produces similar results. [A-1-4; bes1-9638253]

(b) Holonyms

Besides hypernymy, some substitutes for nouns may be in part-whole relation (meronym-holonym in WN terminology) with their respective ft-LUs. In example (6-45), while an author may consider it important to report where the study was conducted, the exact location need not be made explicit in an abstract. In the full text, the place where the study was conducted was given precisely as *the Rocky Mountain Biological Laboratory (RMBL), in the East River valley of the western slope of the Rocky Mountains, near Gothic, Gunnison County, Colorado, United States*. During abstracting, it is out of place to be this precise. While the replacement may be any "smaller part" of the "bigger whole", *Colorado* was chosen as "intermediate" smaller part. Choosing the administrative state to replace the exact location, the research reported is sufficiently situated for readers to relate to the place of study, and without being too precise or too vague. Recall that in example (6-30), the whole was stated without the implicit part.

(6-45) The principal mortality factor for *H. rubicundus* at Gothic was fly predation, especially by satellite flies (*Leucophora* sp.). [D-4-1]

→ *Leucophora* sp. (Diptera: Anthomyiidae) is the principal cause of brood mortality in Colorado. [A-1-4; bes2-9638227]

Holonym(Gothic) = Colorado (¬WN)

Similarly, in example (6-46), it suffice to indicate the season without specifying which month. While much of the general knowledge linking a ft-LU and the substitute used by author is found in WordNet, the problem that remains for a non-author abstractor/system is when to apply such a substitution and what to use as substitute. In example (6-46), while the hypernym for *August* and *September* is *month*, the author chose to use its meronym of *summer*.

(6-46) Hunting spiders alone followed a similar trend, reaching a significantly higher abundance in ant-free trees in late August and September. [R-3-5]

→ The abundance of hunting spiders, the majority being Salticidae, increased significantly by 1.5- to 1.8-fold in trees without ants in the late summer; [A-1-6; oec2-97109313]

Meronym(summer) = August (WN)

While substitutes for general nouns are commonly either Type I (with same stems), or Type III (with hypernyms/holonyms), substitutes for technical nouns are commonly Type III document synonyms, or general words that are in various relations with the ft-LU. Many substitutes also fall in between Type III and IV.

In the following examples, we have provisionally described the relationship between a substitute and its ft-LU as one of part-whole. As much as the meronymic *foliage* in example (6-47) is part of a tree, we may not enumerate the parts constituting a tree. The set is not finite.

(6-47) the biomass of prey in ant-free trees [R-1-2]

→ The biomass of potential prey organisms on foliage, … [A-1-4; oec2-97109313]

Meronym(tree) = leaf (¬ WN; SR)

Syn(leaf) = foliage (WN; SR)

In example (6-48), the parts making up a study is even less definable.

(6-48) In this paper we report on a common garden study of the effects of plant origin and … [D-5-1]

→ Previous common garden experiments also indicated that … [A-1-3; oec1-99120268]

Meronym(study) = experiment (SR);

In example (6-49), while *larvae* may constitute a *brood*, they may not properly be considered as a meronym to *brood* which does not have parts.

(6-49) Our results positively demonstrate that hexane-extractable compounds associated with brood stimulate pollen foraging. [D-5-1]

→ Hexane extracts of larvae containing brood pheromone stimulated pollen foraging. [A-1-4; oec2-97109313]

Mult(larva) = brood (SR);

To carry out Type III substitutions, various kinds of knowledge ranging from world, domain, but especially experimental is required. A substitution is categorized as Type IV, if the link between LUs is not "evident". However, because of a continuum in the range of

substitutions and the different kinds of knowledge involved, we call attention to the subjectivity in categorization for substitutions falling in between Type III and Type IV.

6.2.3.3 Numerical Expressions

Like technical terms and general words, numerical expressions are often reformulated less precisely, although they are no less accurately during abstracting. Substitutions for numerical expressions were observed to be fairly regular. Decimals and absolute figures are rounded off, and percentages replaced by fractions or ratios (see examples below).

(6-50) Fifty-one (16.3%) of the parasitoid species introduced for biological control have been recorded on nontarget native insects. [R-1-1]

→ Sixteen percent of 313 parasitoid species introduced against holometabolous pests are known from natives. [A-1-5; oec1-97112566]

(6-51) The percentage of males among the 585 scored embryos was 17.4, a percentage that was significantly lower than 50%. [R-1-3]

→ Only 17% of 585 embryos sexed from 14 egg sacs were male, a significant departure from a 1:1 sex ratio. [A-1-2; bes1-9946237]

(6-52) a decrease in photosynthetic surface of unpatrolled leaves relative to paired control leaves of about 30-37%. [D-3-5]

→ reduce the leaf area by about one-third. [A-1-2; oec2-97112209]

(6-53) Samples of 300-600 randomly selected woody shrubs [M-1-10]

→ Samples (of up to 600 plants ...) of woody shrubs [A-1-4; oec1-98115427]

While high precision is not important in an abstract, accuracy still needs to be maintained particularly in the context of scientiifc texts. Consider

(6-54) the most active ergatoid male of *C. emeryi* successfully copulated with 36 virgin queens. [D-2-1]

→ Both male morphs are capable of inseminating more than 35 virgin queens. [A-1-8; bes2-9842239]

Ab-LU *35 virgin queens* may not be accepted as substitute for *36 virgin queens*. To maintain accuracy, absolute numbers which have been rounded off, or percentages which have been replaced with ratios or fractions, have simultaneously to be qualified by words expressing vagueness or approximations, such as *more than*, or less commonly *up to*. While it is inaccurate to reformulate *36 virgin queens* as *35 virgin queens*, it is not inaccurate, although vague, to reformulate it as *more than 35 virgin queens*.

If a ft-LU contains a measure noun, then the LU may be substituted with a less precise lexeme that is its holonym or hypernym (see examples (6-55) and (6-56)). To make necessary adjustments to the accompanying cardinal to effect this substitution type, world knowledge is required. Note the large reduction in number of words in example (6-56).

(6-55) For the first 10 weeks, survival ... [R-2-2]

→ survival, ... but after about 3 months, ... [A-1-5; oec2-97112209]

Holonym(week) = month (WN)

(6-56) Beginning 6 weeks after the exclusion of ants, trees were sampled five times at 1-month intervals on 28 June, 30 July, 27 August, 24 September and 29 October 1994. [M-4-6]

→ ... in a 5-month ant-exclusion experiment. [A-1-3; oec2-97109313]

Hypernym(June/July/September/October) = month (WN)

In our study, there were few examples where a substitute is more precise than its ft-LU. One numerical expression was replaced a more exact figure (see example (6-57)), and a quantifier replaced by an exact number (see example (6-58)). Experimental knowledge is involved which is beyond easy access for an abstractor who is not the author.

(6-57) Only three spiders (two lycosids and one salticid) were brought to the nests. [R-9-5]

→ spiders represented only 1.4%[53] of the ants' diet. [A-1-8; oec2-97109313]

[53] The figure was obtained indirectly from information given in table: Tot no. of prey = 216; No. of spiders captured = 3. 3/216 x 100% = 1.4%.

(6-58) Some of the major allelochemicals in tomato are the phenolics rutin and chlorogenic acid and the glycoalkaloid tomatine. [I-5-2]

→ Two of the major allelochemicals in tomato were used: chlorogenic acid and tomatine. [A-1-2; oec1-97109265]

In example (6-59), the generic word was replaced by a more precise substitute. Experimental knowledge was required to effect such a substitution.

(6-59) egg hatch, did not differ between the two species [D-1-5]

→ egg hatch was not significantly different between *O. notulata* and *O. slobodkini* [A-1-7; oec2-97112081]

6.2.4 Type IV

About 1/5 of the substitutions in our corpus are Type IV. Substitutions of this last category are most complicated. There is a continuum of opacity in the link between ft- and ab-LUs as various manipulations and kinds of knowledge are required to arrive at the superordinate hypernymic substitute. Nevertheless, the link is more evident when the context about it is considered (see examples (6-60) to (6-62)).

(6-60) Over four-fifths of the spiders remained on their assigned sites during the first night or longer on daisy [R-4-1

→ However, four-fifths of the individuals that remained a day or longer tended to leave … sooner than daisies [A-1-5; oec1-99120252]

(6-61) no spiders receiving a sole diet of *Drosophila* survived to maturity: all died one or two molts from adulthood. [R-3-7]

→ spiderlings fed solely one of these species did not grow and died without molting. [A-1-7; oec2-97112209]

(6-62) varied significantly among naturally occurring hybrid and parental plants in 1994. [D-1-2]

→ differed significantly among taxa in 1994 and … [A-1-3; oec2-97110360]

Experimental and domain knowledge are required to substitute *diet* with *allelochemical-fed prey* and *growth and consumption* with *performance* in example (6-63). The relation between the LUs is not evident.

(6-63) A three-factor experiment was done to assess the effects of diet$_x$, thermal regime (26: 15°C and 21: 10°C), and gender on predator growth and consumption$_y$. [M-1-1]

→ Two diet experiments addressed the effects of allelochemical-fed prey$_x$ (*Manduca sexta* caterpillars), temperature, and gender on performance$_y$ of the insect predator, … [A-1-1; oec1-97109265]

In example (6-64), domain/world knowledge is required: spiders do not normally forage for their victims, but build webs to capture them.

(6-64) Many spiders of the genus *Argyrodes* (Theridiidae) avoid this cost by foraging on webs of large orb-weaving spiders rather than capturing prey in their own snares. [I-1-3]

→ *Argyrodes antipodianus* is a small kleptobiotic spider that steals prey from webs of the large orb-weaving spider *Nephila plumipes*, … [A-1-1; oec1-97111570]

In example (6-65), some reasoning is required: to cut is to remove manually.

(6-65) We damaged orb webs every 5 days by placing a plastic disc (12.5 cm diameter) against them and cutting a circular section of web around the disc (about 25% of the orb area) with dissecting scissors. [M-7-3]

→ Web loss was evaluated in a separate experiment, by manually removing one-quarter of the web every 5 days for 30 days; … [A-1-7; oec1-97111570]

In examples (6-66) to (6-69) the link between LUs is even more opaque. Experimental knowledge is required to know that the rate of desiccation poses a problem for the individuals having the wounds (in example (6-66)), and that the fruits in example (6-67) were hosts to weevils that oviposited in them.

(6-66) wounds … could exacerbate the rate of desiccation, … [I-2-3]

→ Wounds … might exacerbate this problem. [A-1-2; oec1-98115184]

(6-67) a first class of fruits corresponding either to unsuitable hosts for weevil ovposition or to chestnuts in excess [D-1-1]

→ the first category of hosts includes on average 74% of the chestnuts.

[A-1-9; oec2-98114382]

In example (6-68), the physical location of the abdomen with respect to the body was used as substitute, and in example (6-69), the way in which that particular aspect of the study was conducted was used as substitute.

(6-68) collecting faeces at the abdominal tip [D-7-3]

→ collect their faeces … at the posterior tip. [A-1-3; oec2-99118166]

(6-69) to control for … predilections … to form galls communally. [M-8-2]

→ to compare their … propensities towards communal behavior.

[A-1-4; bes1-9843095]

As with Type III substitutes, some Type IVs substitutes may be semantically about the same, while others are more precise. See examples below.

(6-70) All sperm had been transferred into the seminal vesicles, … [R-11-5]

→ and all sperm is stored in the seminal vesicles. [A-1-5; bes2-9842239]

(6-71) The spiders occur in clumped dispersion patterns near water … [R-11-5]

→ The nocturnal orb-web spider *Larinioides sclopetarius* lives near water …

[A-1-1; bes1-9946043]

(6-72) this response was likely related to prey availability [D-1-3]

→ perhaps in response to prey abundance [A-1-6; oec1-99120252]

(6-73) conspicuous … sexual characteristic … of the wolf spider *Schizocosa ocreata* (Hentz) [I-3-2]

→ wolf spider, *Schizocosa ocreata*, possess a conspicuous … sexual character: …

[A-1-1; bes1-9638017]

6.3 Discussion

6.3.1 Type I

In abstracting, the choice of words to use is no less important than the choice of content to include. We attribute the high percentage of Type I substitutes in abstracting (about 55%), to the need to reformulate content accurately, especially in scientific communications. Reusing the same stem assures a reader that the same concept is being discussed, and that there is no change in topic.

From the regular co-occurrence of certain Type I substitutions with derived forms, we obtain syntactic transformation patterns which provide an abstractor/system with convenient stereotype ways to reformulate content while allowing for shifts in focus and emphasis.

$[A_{modf}\ S_0(V)_{head}]_{NP} \leftrightarrow [Adv_0(A)_{modf}\ V_{head}]_{VP}$ | with undetermined constraints on A and V;

$[A_{modf}\ N_{head}]_{NP} \leftrightarrow [A_0(N)_{modf}\ S_0(A)_{head}]_{NP}$ | with undetermined constraints on A and N;

$[A_{modf}\ N_{head}]_{NP} \leftrightarrow [\ N_{modf}\ S_0(A)_{head}]_{NP}$ | with undetermined constraints on A and N;

What element to focus or emphasize is dependent on a mix of factors such as the "new" context of the abstract, and what is to be communicated. These patterns on Type I substitutions which manipulate words at the local context, involve linguistic knowledge.

Not only is it often not possible to formalize observations in the form of rules, but for the rules identified, we have yet to determine the context under which they might be applied, and the semantic sub-classes involved, although the properties of lexical units are also important. For example, while some transformations, e.g. *differ dramatically* ↔ *dramatic differences*, are bidirectional, others are not possible, e.g. *studied intensely* → **intense study*. We are only just beginning to identify the condensation sub-processes in operation during abstracting, and factors critical on the interplay of these processes still need to be investigated. Nevertheless, we identify the units involved, the knowledge required, the reduction brought about, and partial lists of paraphrases and domain synonyms as partial and short-term solutions to help an abstractor/system in the writing of abstracts.

6.3.2 Type II

Type II substitutes involving all parts of speech, major and minor, were used in the reformulation of content during abstracting. Although no quantitative study was carried out, Type II substitutions appear to involve mainly verbs. Adverbs were substituted most often, and least with nouns. We attribute this to the marginality of the role of modifiers which may be replaced without problem, and the importance of keeping the same stem to refer to the same concept for reasons of accuracy. We speculate that in technical texts, the unnecessary use of synonyms especially with nouns, could signal a false change in focus, or in concept and might confuse a reader. Abstract writing is different from general précis writing where students are told to expressly vary their use of words. While we hypothesize that Type II substitutions might be restricted to lexical units which play a marginal role in the expression of content, this suspicion needs to be verified. Overall, Type II substitutes constitute only a small percentage of substitutions (5%).

As we do not know the reasons for the gains in substitutions with Type II, we are unable to suggest when Type II substitutions should be applied. However, we suspect the reasons to be reader-oriented and related to text revision during writing. In our study, we did not observe any tendency to condense text by replacing longer units with synonymous abstractese forms. Some linguistic units, in fact, had longer substitutes. Meanwhile, we suggest applying Type II substitutions when the units involved are modifiers as observed, or if the substitution brings about an immediate reduction in number of words.

We note that while Type I and Type II substitutions themselves do not bring about obvious reductions in words, each substitute opens new situations for the operation of other condensation processes, or lead to structural changes in the local context as a result of collocational differences.

Consider example (6-22). By substituting *first pair of* with affix *fore-*, not only produces a compact unit, but also allows for more compact reformulation of content. Other linguistic units may also be added on. We make no claims about the operations, their order, or what units are to be attached, but hypothesize how the condensation might have occurred given just the input and output forms.

bristles … on the tibia … of the first pair of legs of mature male *S. ocreata*, …
→ bristles … on the tibia … of the forelegs of mature male *S. ocreata*, …
→ bristles … on the tibia … of **mature male *S. ocreata*'s** forelegs …
→ bristles … on the tibia … of **their** forelegs

Now consider the case where substitution with *fore-* was not carried out. The intermediate forms while acceptable, are more cumbersome, as there are more words and syllables to process.

bristles … on the tibia … of the first pair of legs of mature male *S. ocreata*, …
→ bristles … on the tibia … of **mature male *S. ocreata*'s** first pair of legs …
→ bristles … on the tibia … of **their** first pair of legs …

Consider example (6-14). When adverbial *upon each other* is substituted with adverb *mutually*, the string becomes ungrammatical: ?prey mutually, and local structuring is inevitable.

can potentially compete with and prey upon each other
→ can potentially compete with and ?prey mutually ?text = text is not acceptable;
→ are potential competitors and mutual predators

While WN appears to be an adequate resource for finding synonyms involving lexical words, supplementary lists need to be compiled for function words/expressions and affixes, especially if they involve abstractese forms. Partial lists of different groups of LUs involving domain-related affixes and paraphrases commonly used in biology obtained from the corpus are given Appendix V (see Table A5-2b and Table A5-2c). We did not include abstractese forms and paraphrases for general words and expressions as they can be collated from guidebooks on better writing.

Much of the domain-related information required for substitution was found in WordNet. However, others not found, but which we consider to be potential additions (see Table A5-2a in Appendix V) may be compiled into a special resource. This list of domain-related terms may be seen as an attempt at encoding more stable domain knowledge for the special purpose of abstracting documents on entomology.

6.3.3 Type III

During abstracting, we observed a clear effort by authors to reformulate content less precisely by using less technical terms, less exact numerical expressions, or hypernyms and holonyms which are themselves less precise compared to their respective ft-LUs, or a unit that is in some unpredictable relation with its ft-LU. While the use of less technical terms may be seen as a means to induct new readers into the domain, the use of less precise substitutes in general parallels the reformulation of the gist of a document's content during abstracting. While this is true, tendencies for substitutes to take on more general or more specific meanings compared to their respective ft-LUs are about equal. There seems to be some attempt to equilibrate the change in generality/specificity of meaning in the substitutes is seen in example (6-74), where the increase in generality in a LU is set off by another unit towards more specific.

(6-74) influenced only one aspect of male activity or microhabitat use; [R-3-5]

→ influenced water strider behavior; [A-1-9; oec2-97117258]

Hypernym(activity) = behavior (SR)

In addition to the problem of distinguishing what is linguistic knowledge and what is world knowledge, we draw attention to the subjectivity of what is termed "technical". What makes *arthropod* technical? Its widespread use? Knowledge of reader? Does the fact that many know *deoxyribonucleic acid (DNA)* refers to make *DNA* less technical? While the label in itself is unimportant, it is crucial in the choice of lexical unit to use when abstracting for a particular group of readers.

Because there are just so many general words that may be used as substitutes for technical terms, each scientific sublanguage investigated can have its own innovative mechanism to create domain synonyms: a ft-LU is adjectivalized and used as modifier to a generic word. While possibly less precise, such specially constructed domain synonyms are often more explicit and more susceptible to comprehension because of the generic word. Explicit reformulation of meaning is an important objective during abstracting. Knowledge of such sublanguage devices of forming document synonyms is important in the wording of abstracts.

In abstracts where only the gist of a document's content need be re-conveyed, high accuracy is not required. With numerical expressions, absolute numbers and decimals are consistently rounded off, and percentages replaced by ratios or fractions, before being qualified by words of approximations, e.g. *about*, to maintain accuracy. According to Vande Kopple (1985:84), hedges are validity markers which allow us "to register necessary doubts". Where a numerical expression contains a measure noun, the noun is replaced with one that is less precise, e.g. *week* with *month*, and month *August* with say, season *summer*. According to Meyers (1996:4),

> "vagueness can be used strategically to allow a written text to take on a range of meanings for different audiences with different interests, and to take on new meanings in new situations unforeseen at the time of writing"

Substitution with hypernyms and superordinates may be difficult to formulate in a rule. Choosing a hypernym that is too superordinate, e.g. *animal*, can lead to overgeneralization and unnecessary loss in information, even false statements. Consider

e.g. Vultures began soaring.
→ ?Animals began soaring.

Hypernym(vulture) = bird (WN)
Hypernym(bird) = animal (WN)

Substitution with a hyponym is no less problematic. Not only because of the possibilities to choose from, and the knowledge required to make the choice, but if the substitution itself can be effected without changing the facts. WordNet appears to be an adequate resource for determining the relation between LUs in the corpus studied, even those involving world knowledge[54].

e.g. Vultures began soaring
→ ?Buzzards began soaring Or ?Carrion crows began soaring

Hyponym(vulture) = buzzard, condor, carrion crow, ... (WN)

While experimental and world knowledge are often required to effect Type III substitutions, the tradeoff involved may be large. Depending on the unit used in the replacement, the number of words may be unchanged, or be significantly reduced (see example (6-56)). In view of the possible large reduction in words, this use of less precise forms is worthy of an investigation.

Because of the risk of introducing inaccuracies, Type III substitutions are best restricted to necessary contexts, such as the use of lexical anaphors to avoid repetition, or to opportune situations to insert non-technical units. About 1/5 of substitutions in our corpus are with Type III substitutes.

6.3.4 Type IV

In Type IV substitution, there is no obvious link between ab- and ft-LU. Various kinds of knowledge and manipulation is involved. Despite the high occurrence (20%), we hypothesize Type IV substitution to be attributable to factors such as the inevitable consequence of accommodating the other types of substitution, text composition, target readers, etc..

[54] In general, the ECD does not encode world knowledge. However, see **Gener**(carrot) = vegetable in Melcuk (1995:51). **Gener**(ic) is a lexical function.

6.4 Concluding Remarks

Authors maintain accuracy in abstracting by reusing the stems of ft-LUs in about half of the ab-LUs. Some use of synonyms (Type II substitutes) are observed, but they appear to involve lexical units playing marginal functions, e.g. adjectives and adverbs. During abstracting, authors exhibit an expressed effort to use less technical terms, possibly to help induct newcomers into the domain. Consistent with the objective of abstracts is the less precise reformulation of content. In the scientific corpus studied, numerical expressions are regularly reformulated by rounding off absolute numbers, or using fractions and ratios as replacements. Accuracy is not compromised as they are qualified by words expressing vagueness or approximations.

While we have identified some transformations, and some partial lists of unlikely synonyms: suppletive forms, e.g. *leaf~foliar*, *without~-free*, and general synonyms for technical words, other intensive and comprehensive studies over different areas of biology need to be carried out. Also, a special resource of domain information as given in Table A5-2a in Appendix V needs to be compiled to complement a thesaurus such as WordNet which we found to be impressively rich as a resource for finding related words as replacements.

Besides the above, studies should also be carried out on other technical corpora to determine to what extent formulations identified may be considered as general, and which are domain-related.

Last but not least, studies are needed to determine the knowledge required and the context where each substitution type may be effected, and if Type I substitutions are restricted to important content, and Type II substitutions to marginal content. A reason for the difficulty in the implementation of condensation processes lies in the different kinds of knowledge required, specifically domain and world which requires a different treatment. The interplay of substitution with other condensation sub-processes, and factors linked to purpose and readers at whom the abstracts are targeted are other reasons. We are only beginning to separate out the processes, but still know little about their interplay or how various factors such as communicative intent of author, and readers targeted at, affect substitution.

Chapter 7

Just What may be Deleted, or Added during Abstracting?

Abstracts constituted from extracted sentences are not only disjointed, but also contain unneeded texts. Closely associated with content condensation in the context of summarization is the removal of "delible[55]" information. But just what kind of linguistic units are DELETED during abstracting? And, what functions do these units serve? As seen in section 2.4 and sub-section 5.3.2, the process of condensation may also includes the adding of information by INSERTION or by APPOSITION, depending on whether or not a larger unit is produced by the addition, i.e. information is compacted into an abstract.

By comparing full text sentences used by an author in abstracting with the corresponding sentences in abstract (see Chapter 3 Methodology), this chapter presents a partial inventory of linguistic units that are often ~~deleted~~ (omission is indicated by $\phi_{deleted\_text}$), or added. We discuss the units which may be deleted or added, under various headings: (a) metadiscourse, (b) precision and details, (c) domain, linguistic and experimental knowledge, and (d) explicitness. Just as these units serving various textual functions are deleted, they are also added to provide more disparate information for a compact holistic abstract that is representative of the document. While some types of units are deleted with regularity, others are less predictable. We discuss some of these units below.

7.1 Metadiscourse Units

Metadiscourse is "Writing about writing, whatever does not refer to the subject matter being addressed" (Williams, 1981; cited in Vande Kopple, 1985). Vande Kopple (*ibid.*:83) explains that an author usually writes at two levels. At one level, propositional content on a subject is supplied, and at another, metadiscourse which does not contribute to propositional content, but helps a reader "organize, classify, interpret, evaluate, and react to [the propositional] material", is added. The implication here is that if superfluous metadiscourse in extracted sentences can be reliably identified and deleted, then what is left over is the propositional content, which is the

[55] "Capable of being deleted" (WordNet 1.6).

most important for the abstract. Using Vande Kopple's (1985:83-85) categorization, we discuss some categories of metadiscourse commonly deleted in our scientific corpus.

7.1.1 Illocution Markers

Illocution markers which "make explicit [the] speech or discourse act [being] perform[ed] at certain points" (*ibid.*), concern the relationship between an author and the subject matter, or between author and reader. In our scientific corpus, we identified two types of illocution markers commonly deleted by an author in abstracting. The first type contains first person pronouns *I* and *we* (see examples (7-1) and (7-2)), and to a lesser extent proper names, which we group under the semantic category of AUTHOR (see example (7-3))

(7-1) ~~We found that~~ California gnatcatchers took proportionately more sessile prey than were available in the environment. [D-1-2]

→ $\phi_{illocution\_marker}$ Both adults and young California gnatcatchers consumed more sessile than active prey. [A-1-9; oec1-99120304]

(7-2) ~~I have shown here that~~ there is little support for escape to enemy-free space as a selective factor that maintains this host fidelity. [D-4-2]

→ $\phi_{illocution\_marker}$ There is little evidence to suggest that escape to enemy-free space is a factor that maintains the monophagy of *O. notulata.* [A-1-10; oec2-97112081]

(7-3) ~~Yeargan and Quate demonstrated that~~ juvenile bolas spiders of both sexes attract adult male flies in the genus *Psychoda.* [I-3-6]

→ $\phi_{illocution\_marker}$ Small, early-instar bolas spiders of both sexes attract moth flies in the genus *Psychoda*, ... [A-1-5; oec1-97112572]

A second type of illocution markers involves words such as *test*, *result*, *analysis*, *study*, and *paper*, which can be grouped under separate semantic categories of EXPERIMENT and STUDY, or just STUDY. During condensation, an illocution marker up to and including a complementizer (*that* or *whether*) if present is deleted.

(7-4) ~~Tukey's test showed that~~ the largest reductions in mating activity occurred in the presence of both predators. [R-6-4]

→ $\phi_{illocution\_marker}$ The largest reductions in mating activity occurred in pools with both predators present. [A-1-17; oec2-97117258]

(7-5) ~~The results reported here indicate that~~ these spiders have evolved a specialised foraging behaviour that is tied to the behaviour of nocturnal insects which are attracted to artificial light. [D-7-4]

→ $\phi_{illocution\_marker}$ This orb-web spider seems to have evolved a foraging behaviour that exploits the attraction of insects to artificial lights. [A-1-7; bes1-9946043]

(7-6) ~~A contrast analysis revealed that~~ the mean dominant frequency of responses to the 350-Hz stimulus was significantly lower than the mean dominant frequency of responses to the 450-Hz stimulus. [R-1-3]

→ $\phi_{illocution\_marker}$ In both experiments, males produced calls with significantly lower dominant frequencies in response to each stimulus. [A-1-3; bes2-9945177]

While 2/57 documents in our study excluded first person pronouns from the full text itself, forty percent (22/55) excluded author's overt presence during abstracting.

As much as overt indications of AUTHOR are delible, an author's presence remains implicit in agentless passives (see example (7-7)), and in personifications of inanimate nouns (often from the semantic categories of EXPERIMENT/STUDY) (see example (7-8)), which are typical of scientific writing.

(7-7) In the present study, ~~we investigated~~ the effects of larval shields of *Cassida* spp. that feed upon tansy towards the ant *Myrmica rubra*, … [I-3-1]

→ $\phi_{illocution\_marker}$ The effects of these abdominal shields towards *M. rubra* were studied in three cassidine species, … [A-1-4; oec2-98118166]

(7-8) ~~Based on~~ our observations, ~~we~~ suggest that this is due to interference competition … [I-5-4]

→ $\phi_{illocution\_marker}$ Our observations suggest that … may be due to interference competition … [A-1-10; oec2-97109313]

Also closely associated with these two types of illocution markers are verbs, such as *show*, *suggest*, *reveal*, *indicate*, etc. which we group under the semantic category of COMMUNICATE.

Although mostly deleted, metadiscourse are also added. Some degree of an author's overt or implicit presence is retained in an abstract. Metadiscourse is needed as much in a full text as in an abstract to help a reader interpret the text.

(7-9) We chose the chestnut weevil, *Curculio elephas*, as a model; it is an important pest of the European chestnut, *Castanea sativa*. [M-1-1]

→ We chose the chestnut weevil *Curculio elephas*, a pest of the European chestnut *Castanea sativa*, … [A-1-3; oec2-98114382]

(7-10) Communal gall occupation is related to the density of aphid foundresses on the host plant, and is not necessarily of mutual benefit for gall occupants. [D-1-2]

→ These results suggest that communal gall occupation does not necessarily represent mutual cooperation but may instead be the outcome of competition for limited gall sites on the host plant. [A-1-8; bes1-9843095]

(7-11) Visual observations … together with both remote and onsite atmospheric observations permitted inferences about … [I-4-2]

→ I conducted direct visual observations … concurrently with remote radar observations of aerial plankton … [A-1-1; oec2-99118381]

(7-12) the observations described here strongly suggest that this concentration was "scrubbed" out of the atmosphere by the light precipitation that followed. [D-2-5]

→ I interpreted these observations together with radar data as indicating that (a) large quantities of aerial plankton were entrained by the gust front, "leaked" into the storm outflow, and were subsequently "scrubbed" out of the atmospheric boundary layer by precipitation; … [A-1-7; oec2-99118381]

7.1.2 Text Connectives

Connectives are another type of metadiscourse. These connecting words which link segments of text, lose their function when extracted from the greater context in which they are found. As confirmed in our observations, connectives are almost always deleted during abstracting.

(7-13) ~~However,~~ data on track directions of a large sample of summer gust fronts in east-central Florida indicated no significant orientation. [D-5-1]

→ $\phi_{connective}$ Data on track directions of a large sample of summer gust fronts in east-central Florida suggest that … [A-1-10; oec2-99118381]

(7-14) ~~First,~~ we explore whether the spatial arrangement of male and female embryos within the sacs reflects any possible control of the sequence in which the sexes are laid. [I-4-3]
~~Second,~~ we investigate whether the variance in the number of males produced per clutch may be lower than expected by chance, thus reflecting control of the sex of individual embryos. [I-4-4]

→ $\phi_{connective}$ We also explored the possibility of direct control of the sex of individual offspring in this species by examining the variance in the number of males per sac and $\phi_{connective}$ the spatial distribution of male and female embryos within the sacs. [A-1-3; bes1-9946237]

(7-15) ~~Therefore,~~ the sex of individual embryos can be determined by simply scoring their chromosome number. [I-3-3]

→ $\phi_{connective}$ By scoring the chromosome number of developing embryos, we show that the sex ratio bias of … is the result of an overproduction of female embryos. [A-1-1; bes1-9946237]

7.1.3 Commentaries and Attitude Markers

Commentaries are author's remarks to reader, and attitude markers express author's position with regards to the statement made. Like connectives, these two types of metadiscourse are also almost always deleted. An abstract has little space for direct remarks to reader or to make known author's stand with regards to his findings. However, unlike illocution markers and connectives, these two categories of metadiscourse come in unpredictable forms.

(7-16) small worker size, ~~far from being a handicap,~~ may confer a distinct advantage in these systems. [D-12-2]

→ We propose that the small size of workers $\phi_{\text{commentary}}$ confers a distinct advantage in this system. [A-1-6; oec2-97112209]

(7-17) ~~This transition was somewhat unusual in that~~ convective storms are generally followed by rapid clearing within the study area. [R-2-2]

→ $\phi_{\text{commentary}}$ Clear skies and convective conditions predominated in the area prior to local passage of the gust front. [A-1-2; oec2-99118381]

(7-18) *P. zelicaon* is ~~of particular interest~~, because it is one of the most broadly distributed butterflies ... [I-2-2]

→ *P. zelicaon* $\phi_{\text{commentary}}$ is one of the most widely distributed and polyphagous butterflies ... [A-1-1; oec1-97111209]

(7-19) ~~Whatever the reasons for the unusual concentration in the outflow airmass,~~ the observations described here strongly suggest that this concentration was "scrubbed" out of the atmosphere by the light precipitation that followed. [D-2-5]

→ $\phi_{\text{commentary}}$ I interpreted these observations together with radar data as indicating that (a) large quantities of aerial plankton were ... and were subsequently "scrubbed" out of the atmospheric boundary layer by precipitation; ... [A-1-7; oec2-99118381]

7.2 Precision and Details

In abstracting, where only the core content of a document is important, participial/ relative/subordinate clauses, apposed text, and parenthetical text, which provide precision and details to another unit are more often deleted than added. However, there are other types of units, e.g. quantifiers and determiners, and lexical units which function as nominal attributes, e.g. *aspect* in a nominal complex like *aspect of activity*, which may be deleted or added. In an attempt to compact information into convenient units, lengthy compounds which are typical of scientific texts are often used. However, during abstracting, such compounds may be abridged if its interpretation is not expected to be problematic. We discuss them below.

7.2.1 Elaborating Clauses

During abstracting, clauses providing various details are often deleted. Details may be provided in relative clauses, subordinate clauses, participial clauses. The latter "in comparison with finite subclauses [non-finite clauses] are more economical and avoid repetition; *ing*-clauses and *–ed* clauses, ... [and] are particularly favoured in <formal or written> styles of English" (Leech & Svartvik, 1975:168). Note that such delible clauses are sometimes used to add on information deemed to be important as opposed to the units which were deleted because of their subordinated role.

(7-20) ... 31 populations including 9 from heavy-snowfall regions $_{\text{PARTCL}}$ ... [I-7-1]

→ Thirty-one samples ϕ_{PARTCL} ... [A-1-2; bes1-9946025]

(7-21) In the present study, we investigated the effects of larval shields of *Cassida* spp. that feed upon tansy towards the ant *Myrmica rubra*, a generalist predator. [I-3-1]

→ In the present study, we investigated effects of larval faeces from leaf beetles of the subfamiliy Cassidinae ϕ_{RCL} towards a generalist predator, the ant *Myrmica rubra*. [A-1-2; oec2-99118166]

(7-22) a male secondary sexual characteristic, a conspicuous tuft of bristles and dark pigmentation on the tibia of the first pair of legs of mature male *S. ocreata*, which is lacking in *S. rovneri*$_{RCL}$ [I-5-3]

→ a male secondary sexual characteristic, a conspicuous tuft of bristles and dark pigmentation on the tibia of the first pair of legs of mature male *S. ocreata*, ϕ_{RCL}; [A-1-1; bes1-9638017]

(7-23) In 1995, host feeding predation varied significantly among taxa, but survival and eulophid parasitism did not vary among taxa using Bonferroni criteria. [D-1-4]

→ In the field in 1995, host feeding predation varied significant among taxa ϕ_{but} ϕ_{S}. [A-1-4; oec2-97110360]

(7-24) Sunfish attack water striders from below in deeper water, while fishing spiders perch vertically on rocks and overhanging vegetation along the shore where they may catch and lift water striders off the water's surface. [M-5-2]

→ Green sunfish occupy stream pools and attack water striders from below ϕ_{Conj} ϕ_{S}. [A-1-5; oec2-97117258]

7.2.2 Parenthetical Texts

Parenthetical texts which are deleted, almost always concern information about the experiment. Note in example (7-25) that the "part" in the "part of whole" noun phrase is deleted. Compare this with example (6-45), where the whole is replaced by a part of the whole.

(7-25) Spiders of the nocturnal orb-web species *L. sclopetarius* constructed their webs ~~on the four handrails (length 59 m; height 1.3 m)~~ of the footbridge. [M-1-2]

→ The nocturnal orb-web spider *Larinioides sclopetarius* lives near water and frequently builds webs ϕ $\phi_{Paren\text{-}txt}$ on bridges. [A-1-1; bes1-9946043]

(7-26) A three-factor experiment was done to assess the effects of diet, thermal regime ~~(26: 15°C and 21: 10°C)~~, and gender on predator growth and consumption. [M-1-1]

→ Two diet experiments addressed the effects of allelochemical-fed prey (*Manduca sexta* caterpillars), temperature $\phi_{Paren\text{-}txt}$, and gender on performance of the insect predator, ... [A-1-1; oec1-97109265]

(7-27) The Argentine ant, *Linepithema humile* ~~(formerly *Iridomyrmex humilis*)~~, is one of the most widespread and destructive invasive ants in the world. [I-1-3]

→ The Argentine ant, *Linepithema humile* $\phi_{\text{Paren-txt}}$, has invaded sites across Africa, Australia, Europe, and North America. [A-1-1; oec2-98117420]

In some cases, the parenthetical segments are not deleted, but extracted out into the sentence.

(7-28) cutting ~~a circular section of web around the disc~~ (about 25% of the orb area) with dissecting scissors. [M-7-3]

→ by manually removing one-quarter of the web ... [A-1-7; oec1-97111570]

Because parentheses are convenient ways to add information without disrupting the structure of the sentence, parentheses are also exploited to add disparate information about the experiment or study during abstracting.

(7-29) Gall species were counted by searching entire plants for the presence of insect galls. [M-1-5]

Samples of 300-600 randomly selected woody shrubs were examined at Fynbos sites along transects c. 10 m wide. [M-1-10]

→ Samples (of up to 600 plants per transect for Fynbos) of woody shrubs were investigated for the presence of galls. [A-1-4; oec1-98115427]

(7-30) Contrary to predictions of the temperature hypothesis, *L. dispar* growth rates were higher on foliage from high-elevation tree populations vs. valley tree populations [R-1-1]

Leaf nitrogen concentrations tended to be higher in mountain populations than valley populations for all six tree species, significantly so in pairwise comparisons for five of the species. [R-1-4]

→ Contrary to the temperature hypothesis, high-elevation foliage had higher leaf nitrogen (six of six tree species) and allowed higher growth rates of *Lymantria dispar* larvae (five of six tree species). [A-1-5; oec2-98117133]

In some cases, the transformations may be complicated. Several processes co-occurring. While one bit of parenthetical information is deleted, another bit of information is extracted out of sentence and placed within parentheses as seen in example (7-31): details on *microhabitats* given as within parentheses were deleted. Meanwhile, clause "male water striders shifted to predator-free microhabitats" is nominalized and replaced with hypernym *behavior* to give noun phrase "male water strider behavior", and some explanation given within parentheses.

(7-31) when faced with predation risk in the open water where fish occur, male water striders shifted to predator-free microhabitats ~~(riffles, out of water, edges of pools)~~ and reduced activity that should reduce conspicuousness to predators. [D-2-3]

→ In the presence of both predators, male water strider behavior (microhabitat use and activity) … [A-1-11; oec2-97117258]

7.2.3 Quantifiers and Determiners

During abstracting, quantifiers and determiners which do not contribute significantly to the meaning of the text segment in which they are found, are often omitted. Consider example (7-34). The meaning of the text segment remain generally unchanged with or without the quantifier. Omissions in such contexts are predictable, and can be recovered by a non-naïve domain reader. By such omissions, an author also leave some interpretations open to reader, e.g. if all or some spiderlings dies in example (7-35), thus lending generality to the statements made, and also adds intentional vagueness and reservation to claims made. However, to be equally explicit, an author can also add such linguistic units (see example (7-36)).

(7-32) ~~Several~~ results in this study argue in favour of the hypothesis of host heterogeneity. [D-1-3]

→ ϕ_{Qtf} Our results confirm this host heterogeneity. [A-1-8; oec2-98114382]

(7-33) ~~most~~ prey species face multiple predators [I-1-6]

→ ϕ_{Qtf} prey frequently face multiple species of predators [A-1-2; oec2-97117258]

(7-34) We examine ~~some of these~~ assumptions … [I-1-5]

→ We tested ϕ_{Qtf} ϕ_{Det} assumptions … [A-1-2; bes2-9946171]

(7-35) ~~no~~ spiders receiving a sole diet of *Drosophila* survived to maturity: ~~all~~ died one or two molts from adulthood. [R-3-7]

→ spiderlings fed solely one of these species did not grow and ϕ_{Qtf} died without molting. [A-1-7; oec2-97112209]

(7-36) Cassidine larvae have two movable abdominal spines onto which they collect faeces and/or exuviae with each defecation and moult. [I-2-5]

→ Most cassidine larvae collect their faeces together with exuviae as so-called abdominal defensive shields on two movable spines at the posterior tip. [A-1-3; oec2-99118166]

As with single-word quantifiers just seen, numerical expressions too are rendered vague. In section 6.2.3.3 in the previous chapter, we saw that numbers may be rounded off. However, if the information is considered to be unimportant, it may even be deleted.

As with single-word quantifiers just seen, numerical expressions may be simplified and thus rendered less specific. In section 6.2.3.3 in the previous chapter, we saw that numbers may be rounded off. Numerical expressions may also be deleted entirely in some cases.

(7-37) we excavated all 30_{Qtf} colonies. [M-4-1]

→ we excavated all ϕ_{Qtf} colonies, [A-1-4; bes2-9638145]

(7-38) undertakers were ~~6 times~~ more likely to subsequently remove at least one dead bee [R-3-3]

→ undertakers were ϕ_{Qtf} more likely to subsequently remove a corpse [A-1-3; bes2-9741151]

In others, the numerical expressions may be subjected to complicated transformations (see example (7-40)) which we categorized as Type IV substitutions.

(7-39) high-elevation birch trees had concentrations of condensed tannins as low as half of concentrations in low-elevation trees [D-2-7]

→ high-elevation trees tended to have … lower leaf tannins, … than conspecific trees from lower elevations [A-1-6; oec2-98117133]

7.2.4 Nouns Providing Precision and Detail and Attributes of Nouns

Besides quantifiers and determiners, lexical units providing precision or detail to a noun are also often deleted. The deleted unit may be an aspect or attribute to a noun, or its nominal complement. In example (7-40), *aspect* is an attribute of the noun *activity*. Deleting the attribute, removes the precision on the noun. However, it is not the attribute that is always dispensable. In example (7-41), it is the nominal complement *life history of this annual species* which provides details about the head noun *study* that is deleted.

(7-40) influenced ~~only one aspect of~~ male activity or microhabitat use; [R-3-5]

→ influenced ϕ water strider behavior; [A-1-9; oec2-97117258]

(7-41) A recent study ~~of the life history of this annual species~~ revealed an unusually extended reproductive period, which results in a very wide and possibly bimodal size distribution of the coexisting juvenile instars. [I-6-2]

→ Preliminary field observations ϕ indicated an extended reproductive period, which results in a very wide size distribution of juvenile instars. [A-1-3; bes1-9945349]

In example (7-40), the first noun or attribute is the one that is dispensable with regards to the verb. Deletion of the second noun leads to semantic incompleteness or ill-formedness.

**influenced only one aspect* ~~*male activity*~~

influenced ~~*only one aspect*~~ *male activity.*

In example (7-41), it is the complements which are dispensable. Deletion of the first noun leads to ungrammaticality.

A recent study ~~*of the life history of this annual species*~~ *revealed an extended reproductive period*

*~~*A recent study of*~~ *the life history* ~~*of this annual species*~~ *revealed an extended reproductive period*

*~~*A recent study of the life history of*~~ *this annual species revealed an extended reproductive period*

In the above, it is unfortunate that deletion is not associated with the position of the noun in the $NOUN_1$-of-$NOUN_2$ construct, but with its importance for which a greater local context involving the verb has to be considered. Sinclair (1991:81-98) discussed the problem of determining which of the noun may be the head, and highlighted the case of what he calls "double-headed nominal groups" where both nouns are equally important as heads, e.g. *the growth of a single-celled creature*, *the design of nuclear weapons*.

While deletion is not linked to position of noun in $NOUN_1$-of-$NOUN_2$ construct, neither does it appear to be word-dependent except for *species* which is deleted in two clear situations: (a) if the meaning of phrase *species of X*, is the same as *X* itself (see example (7-42)), and (b) if it is clear that the noun in question is a species name, i.e. *R. alternata* (see example (7-43)).

(7-42) specializes on a few ~~species of~~ moths [I-2-2]

→ attract certain ϕ male moths [A-1-3; oec1-97112572]

(7-43) Many ~~species of~~ the tephritid genus *Rhagoletis* are very common. [I-2-2]

→ *Rhagoletis alternata* is a common ϕ tephritid fly [A-1-1; oec1-98115154]

We note that not all lexical units functioning as an attribute are consistently deleted. *Predictions* which was deleted in example (7-44), was added in example (7-45) even though it was redundant.

(7-44) Contrary to ~~predictions of~~ the temperature hypothesis, [R-1-1]

→ Contrary to ϕ the temperature hypothesis, [A-1-5; oec2-98117133]

7.2.5 Compound Nouns

In compound nouns, modifiers which have been previously used with, or linked, to its head, may be deleted in text development. In example (7-45), the modifier *pollen* which was linked to *foraging* in preceding ab-sentences was deleted without detrimental loss in meaning. Linked modifiers in intervening positions in compound are especially omissible. In example (7-46), *nymphal population* was mentioned on two previous occasions.

(7-45) to test the two ~~pollen~~ foraging-regulation hypotheses: [I-5-4]

→ test the predictions of two ϕ foraging-regulation hypotheses: ...

[A-1-3; bes2-9844193]

(7-46) Nymphal ~~population-level~~ diet breadths (H) were less than adult diet breadths at all localities. [R-5-1]

→ Nymphal ϕ diet breadths were significantly less than adult diet breadths at four of six localities and … [A-1-5; oec2-99120437]

7.3 Domain, Linguistic and Experimental Knowledge

Deletion may be applied for reasons of redundancy, or because the information is felt to be implicit to reader. To provide disparate information about the objects studied or the experiment, units are often inserted, as modifiers and sentential adverbials, or in apposition. At the same time, a modifier containing information that is discoverable from the head noun or recoverable from context, may be deleted. We look at some of them below according to the knowledge involved.

7.3.1 Domain Knowledge

While explicit mention of the social nature of spider mites in example (7-47) may be informative in the full text to novice readers, it is not crucial at the point of reading abstract to determine pertinence of document, and hence may be deleted. Similarly in example (7-48), it suffice just to know that the document in question concerns tomatine. In both sentences, the deleted information is a classificatory detail. The spider mite is social organism, and tomatine is a glycoalkaloid. We note in passing the use of the punctuation colon to present content in "telegraphic" style.

(7-47) In a ~~subsocial~~ spider mite, *Schizotetranychus miscanthi* Saito, … [I-2-1]

→ the ϕ spider mite, *Schizotetranychus miscanthi* … [A-1-1; bes1-9946025]

(7-48) Some of the major allelochemicals in tomato are the phenolics rutin and chlorogenic acid and the glycoalkaloid tomatine. [I-5-2]

→ Two of the major allelochemicals in tomato were used: chlorogenic acid and ϕ_{modf} tomatine. [A-1-2; oec1-97109265]

7.3.2 Linguistic Knowledge

In example (7-49), the insertion of modifier *young* is superfluous. The semantic component of *immature* is implicit in *larvae*. While this is true, and WordNet may give *immature* as a synonym of *young*, a way has to be found to detect the redundancy in *young* and *larva*.

(7-49) The presence of ~~young~~ larvae also affects the proportion of foragers collecting pollen:... [I-2-5]

→ The decision to collect pollen by honey bee foragers depends on the number of ϕ larvae (brood), ... [A-1-2; bes2-9844193]

Syn(young) = immature (WN)

larva = immature free-living form (WN)

7.3.3 Experimental Knowledge

7.3.3.1 Modifiers

In scientific reportage, it is implicit that observations must be *significant* to be reported, and that assessments when made are *relative* to something else. While implicit, the evaluative modifier of *significant* was retained 2-3 times more often than they are deleted (see Table A6-2 in Appendix VI). *Significant* is an important lexical unit to include to add credibility to claim made. Compared to modifier *significant*, there were not as many occurrences of *relative* in our corpus.

(7-50) spiders ~~significantly~~ influenced only one aspect of male activity or microhabitat use; [R-3-5]

→ Spiders also ϕ influenced water strider behavior; [A-1-9; oec2-97117258]

(7-51) male frogs use call frequency, ... to assess the ~~relative~~ size of other males [I-3-2]

→ the ability of male green frogs to assess the ϕ size of an opponent [A-1-1; bes2-9945177]

Other less commonly used evaluative modifiers, e.g. *dramatic* and *readily* which are not implicit are almost always omitted.

(7-52) a ~~dramatic~~ stepwise change in behavior around a set point [D-3-5]

→ a ϕ stepwise change in foraging activity as pollen storage levels moved beyond a set point. [A-1-4; bes2-9946171]

(7-53) that can be ~~readily~~ applied to address these questions … [I-4-5]

→ that can be ϕ applied to test for departures from … [A-1-8; bes1-9946237]

7.3.3.2 Adverbials

Like modifiers, adverbials are marginal in their function and are often deleted. The adverbials deleted in our corpus commonly provide details about the experiment. However, they may also be added as in example (7-55) to provide disparate information about the experiment.

(7-54) Sunfish attack water striders from below ~~in deeper water~~, … [M-5-2]

→ Green sunfish occupy stream pools and attack water striders from below $\phi_{adverbial}$. [A-1-5; oec2-97117258]

(7-55) both having the same clinal trend ~~in the relationship between male aggressiveness and relatedness created by winter coldness~~. [D-3-5]

→ each having a similar clinal trend $\phi_{adverbial}$ within Japan. [A-1-4; bes1-9946025]

7.4 Explicitness

7.4.1 Hypernym

A common noun, scientific name, or a technical name may be made more explicit by postposing it with its hypernym which contains redundant information. A hypernym was not found to be deleted, when it follows a technical name, e.g. *chrysomelid beetles*.

(7-56) the host ranges of two ~~species of~~ chrysomelid beetles, ~~*Ophraella notulata* and *O.slobodkini*~~ that are specialized on different species in the Asteraceae. [I-6-1]

→ in the host specialization of two chrysomelid beetles ϕ that are monophagous on different species of Asteraceae. [A-1-1; oec2- 97112081]

Hypernym(chrysomelid) = beetle (WN)

However, when it follows a scientific name (see example (7-57)), it is deleted about half of the time; in half of these cases, the hypernym involved is the generic *species* or *genus*.

(7-57) there was a … decline in the density of the mature *S.punicea* ~~plants~~ … [R-2-2]

→ There has been a … decline in the density of mature *S.punicea* ϕ …

[A-1-6; oec1-98114343]

Hypernym(*S. punicea*) = plant (SR)

(7-58) *Pseudacteon* parasitoids were active during daylight ~~hours~~ at temperatures above 18°C. [R-4-1]

→ *Pseudacteon* parasitoids commonly attacked Argentine ants, but not other ant species, in daylight at temperatures above 18°C. [A-1-5; oec2-98117420]

Hypernyms postposed to lexically ambiguous common nouns are not deleted (see example (7-59)), except in the case of partial repetitions (see example (7-60)). Where absent, an appropriate one may even be inserted. A common domain-related hypernym for an ACTIVITY noun is *behavior*.

(7-59) showed a significant change in foraging activity, … [R-1-1]

→ showed a stepwise change in foraging activity as … [A-1-4; bes2-9946171]

Hypernym(foraging) = activity (WN)

(7-60) whether Argentine ant foraging in Brazil is suppressed by … [I-4-1]

→ the foraging behavior of Argentine ants … in southern Brazil. …

[A-1-4; oec2-98117420]

Hypernym(foraging) = behavior (¬WN)

In example (7-61), *prey organisms* which was given in full form in a preceding ab-sentence, was partially repeated in subsequent mentions with-out its hypernym. Hence, while explicitness is important, deletion is also conditional on other lexical units in the abstract.

(7-61) The most abundant prey ~~organisms~~ brought to the nest were Aphidoidea … [R-9-4]

→ The majority of prey ϕ captured by ants were Aphidoidea … [A-1-8; oec2-97109313]

Hypernym(prey) = life form, organism, … (WN)

7.4.2 Emphatic *both*

As with lexical words, emphatic redundant function words, e.g. *both*, too may be deleted (see example (7-62)), or retained (see example (7-63)).

(7-62) ~~Both~~ undertakers and guards were less likely to engage in behavior typical of young bees than food storers and wax workers. [D-3-2]

→ ϕ_{Qtf} Guards and undertakers were less likely to perform behavior normally associated with young bees compared to food storers and wax workers. [A-1-6; bes2-9741151]

(7-63) The presence of fish caused decreases in both mating frequency and mating duration, while spiders caused a significant reduction in mating duration, but not mating frequency. [R-6-3]

→ The presence of fish reduced both the number of matings per pool (mating frequency), and mean mating durations. [A-1-15; oec2-97117258]

7.5 Discussion

7.5.1 Metadiscourse

As much as metadiscourse does not contribute to propositional content, not all types of metadiscourse are deleted to the same degree during abstracting. While connectives are almost always deleted, three-fifths of abstracts contain overt indications of author's presence. According to Meyers (1992:297), metadiscourse phrases such as *We found that*, are stereotypical means of making "strong, distinctive, but polite claims". Such phrases, while superfluous, serve to make claims, and are important if an author wants his findings to be accepted. Also, while some markers are more characteristic and detectable as they involve special constructs, certain semantic categories of words, or are from a definable set of words, others such as commentaries, are less predictable.

Given that overt indicators of AUTHOR commonly used are mainly *we* and *I*, and the list of connectives is identifiable and finite, these two types of metadiscourse may be deleted from extracted sentences without any computation of salience, which is desirable for automatic summarizers. However, the text left over after metadiscourse deletion is rarely used in its entirety, but is often further subjected to other condensation processes. For this, a study into the

manipulation of leftover text to accommodate resulting structural and lexical changes, and its interplay with other condensation sub-processes, is necessary. Less than two percent of absentences in our corpus were simple direct extracts (see example (7-64)).

(7-64) ~~Finally, they show that~~ the comb itself, rather than the brood within it, is sufficient to produce the negative feedback, although the brood may also contribute to the effect.

[D-1-3]

→ $\phi_{connective}$ $\phi_{illocution\_marker}$ The comb itself, rather than the brood within it, is sufficient to provide the negative feedback, although the brood may also contribute to the effect.

[A-1-6; bes1-9842193]

From the illocution markers deleted during abstracting, we note that an explicatory statement of what was carried out, or observed in a study may be reformulated using any in a continuum of abridged forms (see section 7.1.1). Consider a meta-construct of semantic categories such as: [AUTHOR COMMUNICATE X [PREP EXPERIMENT/STUDY]$_{adverbial}$]$_S$, where AUTHOR is a first person pronoun such as *I* or *we*, COMMUNICATE can be a verb such as *show* or *indicate*, and EXPERIMENT/STUDY can be a noun such as *result* or *paper*. By deleting AUTHOR and personifying STUDY in sentential adverbial, the original construct S is condensed to S': [STUDY COMMUNICATE X]$_{S'}$. Intermediate construct S' may in turn be condensed to construct S" by deletion and passivization: [X BE COMMUNICATE]$_{S"}$, and sentence S" may still further be condensed to just X (see below).

[~~AUTHOR~~ COMMUNICATE X [PREP STUDY]$_{adverbial}$]$_S$

→ [~~STUDY~~ COMMUNICATE X]$_{S'}$

→ [X ~~BE COMMUNICATED (BY STUDY)~~]$_{S"}$

→ [X ~~BE COMMUNICATED~~]$_{S"'}$

→ [X]$_{S"''}$

Note that while we make no claims on the order, or the processes involved, we suggest how the final and intermediate constructs might have been obtained. Any one of these constructs may, at any stage, be subjected to other condensation sub-processes. The interplay of processes in content condensation is an area which requires investigation.

In his abstracting procedure, Saggion (2000:57-58) exploited nouns from the semantic categories of AUTHOR and STUDY/EXPERIMENT which he termed DOMAIN CONCEPTS, and verbs from the category of COMMUNICATE, which he referred to as DOMAIN VERBS, to identify sentences for producing different kinds of abstracts. The selected sentences were then "rephrased" using various transformations (see Table 2-5 in sub-section 2.4.6). Saggion (*ibid.*) was, in effect, using illocution markers to identify sentences, before applying the process of deletion.

7.5.2 Precision and Details

The restricted space of an abstract has little place for precision and details. While parenthetical and apposed texts are most reliably deleted, other linguistic units serving the same function may for various reasons also be retained or even inserted, as often happens with experimental details to pack in some background information.

Critical for the deletion of measure nouns is the significance of its contribution in content. Often, to smooth "awkward" co-occurrence of words as a result of deletion, and to compensate for loss in precision, other condensation sub-processes are involved.

While repeated use of word within a text is expressly avoided, repetitions consequent from deletions are often removed. In example (7-56), the lexical unit *species* which was one too many after deletion of apposed text, was removed despite the ambiguity in reading it creates for a naïve reader who may interpret *two chrysomelid beetles* as "two individuals of chrysomelid beetles". The deletion is indicative that abstract is target-reader sensitive.

In NOUN_1-of- NOUN_2 constructions, deletion of a noun appears to be related to its relative importance in the structure governed by local context, i.e. its dispensability[56]. Where either noun may be deleted, it suffices as is often the case, to keep the noun that expresses the "gist". Where the first noun is *species*, NOUN_1may be deleted if NOUN_2 alone conveys the meaning in NOUN_1-of-NOUN_2, or if it is clear from context that the noun referred to in the local context is a species name.

[56] For a discussion of the problem of interpreting N-of-N constructions see Sinclair (1991:81-98).

7.5.3 Redundancy, Emphasis and Implicitness

Despite the redundancy it introduces, hypernyms apposed to a noun are most of the time retained. While there are clear situations where they may be deleted, the reason for which they are retained appears to be varied.

In the case of a technical noun, retaining a hypernym makes explicit the unfamiliar meaning that the technical noun carries. While a scientific name may also be said to be technical, it is clear to a novice reader that it is a proper name. With a technical name, it is not clear to a novice reader what the noun refers to, and this hampers understanding. In example (7-56), while a novice may not know what *chrysomelids* are, it suffices in an abstract to know that they are beetles. Contrary to technical nouns which usually have but one restricted meaning, common nouns may be polysemic or lexically ambiguous. Keeping its hypernym, or inserting an appropriate one for the context where it is absent, delimits and makes explicit the meaning of the common noun. Scientific documents cannot be ambiguous. The non-deletion and insertion of hypernyms appears to be reflective of the readership at which an abstract is directed, and of author's effort to make explicit unfamiliar knowledge to novice readers.

WordNet appears to be an adequate resource for finding relations between words, even those involving knowledge in biology, e.g.

Hypernym(*Opuntia*) = cactus (WN),
Hypernym(*Cecropia*) = dicot genus (WN),
Cecropia = tropical American trees (WN)

A common noun may even be linked to its scientific name, and there were very few words whose relation were not found in WN. However, a problem that confronts abstracting by sentence extraction is in deciding which among the possible hypernyms to use. Some words may have their own domain-related hypernym which are unlikely to be found in a general-purpose thesaurus such WN. For this, a special thesaurus supplemented with domain knowledge will be invaluable.

Content selected for inclusion in an abstract may be presumed to be important, and hence, does not require further emphasis. Hence, it is probable that units deleted are redundant lexical modifiers and function words which serve an emphatic function. However, to draw attention to important points, especially if the material is unfamiliar, they may equally be retained.

Also, while findings on the deletion of evaluative modifiers are not conclusive, modifiers *readily*, *dramatic* appear to be almost always deleted. Changes need not be dramatic, and actions can be carried out other than readily. However, in scientific reportage it is necessary that observations be objectively conducted, i.e. relative to some measure, and to be significant to be reported. Modifiers *significant* and *relative* which are retained more often than they are deleted, lend confidence to their credence.

7.6 Concluding Remarks

Linguistic units commonly deleted include: illocution markers containing first person pronouns, connectives, parenthetical texts, apposed texts and repetitions. While deletion of such linguistic units may be a first step in condensation, multiple deletions of such units alone can significantly abridge a text without critical loss in core content.

(7-65) although ~~high~~$_x$ temperatures ~~clearly~~$_y$ had a ~~suppressive~~$_z$ effect ~~on foraging~~$_w$. [R-3-3]

→ although ϕ_x temperature ϕ_y had some ϕ_z effect ϕ_w. [A-1-10; oec2-98117420]

As much as example (7-65) seems direct, the situation is usually much more complicated as seen in example (7-66),

(7-66) ~~We used~~ a video imaging technique ~~as another way~~ to test ~~the hypothesis that~~ asymmetry in tufts of male *S. ocreata* influences female receptivity, as it allows manipulation of the presence or absence of tufts while controlling for behavioral differences ~~among males~~. [M-6-3]

→ As a test for concomitant behavioral effects, female spiders were shown video images of a courting male with symmetric tufts and the same video image altered to have asymmetric tufts. [A-1-5; bes1-9638253]

where various manipulations and transformations, and decisions requiring knowledge, are required to get to the final ab-sentence. To effect the following transformation: *to use an imaging technique to test X → to alter an image to have X*, an abstractor needs to know that one can

"use" an imaging technique by "altering" an image, and that "female receptivity" refers to a female's response to courting males, and not her receptivity to some other behavior, say feeding.

While an investigation into factors critical for reliable identification of dispensable units will be invaluable to condensation, other studies should include: (a) the extent to which indication of an author's presence, overt and non-overt, may be deleted, (b) the effect of deletion on readability, and (c) deletion in NOUN_1-of-NOUN_2 constructions according to semantic classes of noun, or type of noun, in both general and other domains. Pending such long-term studies, short-term projects can concentrate on finding inventories for different groups of linguistic units, that may be deleted to complement summarization by sentence extraction (see Table A6-1 in Appendix VI for some examples of illocution markers deleted). A list of nouns for which *behavior* is its hypernym will be useful in our domain.

To advance research on summarization which has largely been limited to selection, future studies must be carried out to determine other sub-processes in condensation, and how the processes interact with each other, and with linguistic manipulations implicated.

Chapter 8

Just What may be Compressed in Abstracting?

In sub-section 5.3.3 we defined compression as the re-expression of a linguistic unit in fewer essential lexemes, and in Chapter 6 we considered some replacements to be the result of compression of some linguistic units into fewer essential units which are then used in substitution. Chapter 8 will look some complex linguistic units which are often condensed into compressed forms when abstracting scientific documents on biology, in two sections: (a) compression of verbal complexes/phrases and (b) compression of nominal complexes. The linguistic units compressed are **underscored in bold**.

8.1 Compression of Verbal Complexes/Phrases

Some linguistic units involving verbal complexes were compressed during abstracting include: (a) complexes containing support verb, (b) complexes containing catenative, and (c) phrasal verbs. Expansion of verbal complexes was not evident in our study.

8.1.1 Complexes with a Support Verb: $V_{support} + S_0(V_x) \rightarrow V_x$

In English, if the verb in a [V + N] collocation may be deleted, and the noun N verbalized, and used to replace the whole collocation and with the meaning remaining essentially unchanged, then the verb V is said to be semantically empty, and is a light, or support verb[57].

$V_{support} + S_0(V) \rightarrow V$ [58]

Meaning-Text Theory (MTT) has lexical functions, e.g. $\mathbf{Oper}_1$(blow) = [to] deal [ART ~ to N] (see Melcuk, 1996:61), to describe the relation between such a support verb and the noun. While the preferred way to describe the operation is to use deep-structure paraphrase in MTT, one could describe it purely on the surface level as we have done. Our aim is not to generate but to illustrate the operations.

[57] Melcuk (1996:60) calls them *semi-auxiliaries*.

[58] Using ECD notation.

In most of the examples in our study, the noun is a derived form. Hence, the process that follows in not one of verbalizing the noun, but de-nominalizing a derived noun. The meaning of the verbal complex is concisely retained in the verb, i.e. the process is one of compression. Note that in the examples given below the operations are accompanied by other processes such as passivisation during abstracting. In example (8-1), lexical unit *pronounced* was replaced by the more precise expression of *by a factor of 3.2*. Studies into the interplay of various processes, condensation and non-condensation are much needed.

(8-1) The observation that fluctuations of CO_2 concentration around a plant lead to a pronounced reduction of oviposition indicates that [D-1-1]

→ On host plants exposed to rapid fluctuations in CO_2 concentration, the frequency of oviposition was reduced by a factor of 3.2. [A-1-4; oec2-97110539]

(8-2) The complex leaf litter habitat of *S. ocreata* may create an important physical constraint on the effectiveness of vibrational signalling; [D-5-2]

→ as vibratory communiction is constrained by the complex leaf litter habitat of some populations. [A-1-8; bes1-9638017]

In example (8-3), the verb is not a simple verb but involves a derived adjective.

(8-3) sensory organs that are specialised to the detection of CO_2 find their strongest expression in … herbivorous Lepidoptera. [I-1-7]

→ Sensory organs that detect CO_2 are common in herbivorous moths and butterflies, … [A-1-1; oec2-97110539]

We hypothesized the transformation to be as follows. First, the source linguistic unit is compressed. Next, the substitute is optionally replaced with a synonym. In example (8-3), the substitute *detect* was not further replaced with a synonym. In example (8-4), linguistic unit *caused decreases* is compressed to lexical unit *decreased*, before being further replaced. *Decreased* was replaced by *reduced* which is its troponym of *decrease* in WN (see sense 2). With verbs, the substitute is very often a troponym to the unit replaced.

(8-4) The presence of fish **caused decreases** in both mating frequency and mating duration, ... [R-6-3]

→ The presence of fish ϕ **reduced** both the number of matings ... and mean mating durations. [A-1-15; oec2-97117258]

Troponym(decrease$_V$) = reduce$_V$ (WN)

In example (8-5), while *argue in favour of* can be reduced to *favour*$_V$, deletion is better followed by substitution with *confirm*. Acceptability of form left over following deletion of other units is questionable (indicated by ?).

? ϕ results ϕ ϕ favour$_V$ ϕ host heterogeneity

results confirm$_V$ host heterogeneity

(8-5) ~~Several~~ results ~~in this study~~ **argue in favour of** ~~the hypothesis of~~ host heterogeneity. [D-1-3]

→ ϕ Our results ϕ ϕ **confirm** this ϕ host heterogeneity. [A-1-8; oec1-99120252]

Hypernym(favour$_V$) = permit$_V$ (WN)

Hypernym(confirm$_V$) = permit$_V$ (WN)

In example (8-6), deletion is accompanied by substitution because of unacceptability (indicated by *) of lexical form.

**a factor impacting population size*

a factor influencing population size

(8-6) Cannibalism can also **have an impact on** the size structure of populations. [I-2-5]

→ Cannibalistic tendencies …may be a significant factor ϕ **influencing** population size. [A-1-1; bes1-9945349]

Troponym(impact$_V$) = influence$_V$ (WN)

This particular compression process may be seen as a special case of Type I substitution.

8.1.2 Complexes with a Catenative: CATENATIVE + VERB$_{non\text{-}finite}$ → VERB

A catenative is "a lexical verb which governs the non-finite form of another lexical verb" (Crystal, 1997). During abstracting, a catenative may in few and restricted cases be replaced by the non-finite verb if its deletion does not bring about a change in meaning. It is not always possible to know linguistically if the change in meaning is marginal, e.g.

'X was allowed to hatch' ≅ 'X hatched',

but 'X stridulate to support Y' ≠ 'X support Y'.

In some, as seen from examples (8-7) and (8-8), only the author-researcher knows if a catenative may be deleted.

(8-7) foragers **tends to move** a small distance. [I-1-2]

→ naive bees … ϕ **flew** shorter distances … [A-1-7; bes2-9639381]

(8-8) high-elevation foliage **tended to support** higher first instar growth than … [R-9-3]

→ high-elevation trees **tended to** … **support** higher insect growth performance than … [A-1-6; oec2-98117133]

In some examples, the transformation is a bit more complex with the interplay of other condensation sub-processes. In example (8-9), after compression from *appear to depend on*, the

compressed form *depend on* is further replaced with *affected*. We admit that the process may arguably be seen as a single substitution with *affected*.

(8-9) However, the outcome of pairwise male-male interactions does not appear to depend on body size or balloon size despite frequent and vigorous male-male interactions. [D-1-2]

→ We found that neither male body size nor balloon size affected the outcome of pairwise male-male interactions. [A-1-7; bes1-9945161]

8.1.3 Prepositional Verbs: VERB + PREP → PREP

A prepositional verb is a complex of a verb and a preposition[59]. In example (8-10), while the verb is deleted from the VERB+PREPOSITION complex, we hypothesize the transformation to be more than just a simple deletion. We prefer to see the process as a compression of meaning of the verb which is now implicit in the preposition. This is an interesting area of study to look into in future work.

(8-10) Plants either of whose parents originated from the Bayshore location ... than ... [D-1-3]

→ Plants with parents from one of three locations ... [A-1-5; oec1-99120268]

(8-11) the CO_2 gradients that normally occur in the vicinity of a plant are essential key stimuli within the context of oviposition ... [D-1-1]

→ As the CO_2 gradients $\phi_{verb\text{-}prep}$ in the vicinity of a host plant depend on its physiological condition, [A-1-2; oec2- 97110539]

8.2 Compression/Expansion of Clauses

8.2.1 Nominalization/De-nominalization

8.2.1.1 Nominalization

We restrict our use of nominalization to that of derivation of a noun phrase from an underlying clause. Nominalization is a special case of compression.

(8-12) Many species can potentially $compete_V$ with and $prey_V$ upon each other. [I-1-2]

→ Spiders and ants are potential $competitors_N$ and mutual $predators_N$. [A-1-1; oec2-97109313]

---

[59] To distinguish between a phrasal verb and a prepositional verb, see Leech & Svatvik (1975:264-265) who gave four differences.

(8-13) females mated with the first male to court$_V$. [R-1-3]

→ females mated more often with males that initiated courtship$_N$ first,

[A-1-4; bes1-9638017]

8.2.1.2 De-nominalization

During abstracting, noun phrases were sometimes de-nominalized.

(8-14) pronounced reduction$_N$ of oviposition [D-1-1]

→ oviposition was reduced$_V$ by a factor of 3.2 [A-1-4; oec2- 97110539]

(8-15) highly significant differences$_N$ among plants in survival [R-4-1]

→ *P. salicifoliella* survival differed$_V$ significantly among three willow taxa

[A-1-2; oec2-97110360]

De-nominalization may be accompanied by substitution.

(8-16) additional ovipositions$_N$ are not thought to occur in error. [I-3-1]

→ to lay$_V$ additional eggs should therefore be based on ... [A-1-2; bes1-9639061]

8.2.2 Personification/De-personification

8.2.2.1 Personification

In sub-section 7.5.1, we saw the personification of inanimate entities. Personification is a kind of compression. The presence of author is set into the background.

(8-17) Here we$_{human}$ examined whether Argentine ant foraging in Brazil is suppressed by the presence of Pseudacteon parasitoids [I-4-1]

→ This study$_{inanim}$ examined the effects of parasitoid flies, genus Pseudacteon, on the foraging behavior of Argentine ants in part of their native range in southern Brazil.

[A-1-4; oec2-98117420]

(8-18) Here, I$_{human}$ will investigate the question of whether predation and parasitism play an important role in ... [I-6-1]

→ This paper$_{inanim}$ examines the role of predation and parasitism in the host specialization of two chrysomelid beetles [A-1-1; oec2- 97112081]

$[N_{human}\ V\ X]_S \rightarrow [N_{inanim}\ V\ X]_S$

8.2.2.2 De-personification

As with de-nominalization, de-personification sometimes is applied to make explicit the presence of the author. Studies into the context of personification and de-personification are required.

(8-19) This study inanim demonstrates that small, "unaggressive" plant-ants can be quite effective anti-herbivore defenders of their host plant. [D-12-1]

→ In this study, we human demonstrate that an important benefit provided by the small host-specific ant *Petalomyrmex phylax* to its host plant *Leonardoxa africana* is efficient protection against herbivores. [A-1-1; oec2-97112209]

8.3 Compression/Expansion Involving Nominal Complexes

8.3.1 Compression to Compound Noun by Deletion

Besides verbal phrases and clauses, other groups of words may also be compressed. Complex nominal complexes may be compacted to compound nouns.

(8-20) indicator of male quality [D-1-4]

→ quality indicator [A-1-7; bes1-9638253]

(8-21) Argentine ants avoided high rates of parasitoid attack. [R-6-3]

→ Parasitoid attack rates diminished as Argentine ants retreated underground. [A-1-7; oec2-98117420]

(8-22) most of the recorded species are oligophagous, feeding on more than two genera within the Brassicaceae. [R-2-1]

→ species are predominantly oligophagous, feeding on more than two Brassicaceae genera. [A-1-4; oec2-98113391]

(8-23) the ecology of the ant in South America requires further study. [I-3-5]

→ the ecology of Argentine ants in their native habitat. [A-1-3; oec2-98117420]

(8-24) the handrails were equipped with artificial light tubes at … [M-1-3]

→ the artificially lit handrails [A-1-2; bes1-9946043]

Often, the process is additionally accompanied by the adding on of other units.

(8-25) occurred in the presence of both predators. [R-6-4]

→ occurred in pools with both predators present. [A-1-17; oec2-97117258]

(8-26) small worker size, [D-12-2]

→ small size of workers [A-1-6; oec2-97112209]

8.3.2 Expansion of Noun Phrases to Complex Noun Phrases

Expansion of compound with expansion of substitution of a more explicit form.

(8-27) [chemical analyses $_{head}$]$_{NP}$ [I-2-3]

→ [analyses $_{head}$ [of foliar nitrogen and condensed tannin]$_{PP}$]$_{NP}$

[A-1-4; oec2-98117133]

(8-28) the female response to [shaved males]$_{NP}$ was significantly lower … [R-4-3]

→ females showed receptivity less often to [males $_{head}$ [with tufts removed]$_{PP}$]$_{NP}$.

[A-1-6; bes1-9638017]

8.4 Semantic Compression

Under substitution we identified a group of replacements which are in some ways the result of compression process. Multiple units are replaced by fewer units. However, because of the interplay of other transformations, the resulting unit is in conflict with our definition of compression (see example (8-30) to example (8-32)). The unit replacing the unit from full text may not share lexeme (see example (8-30) and example (8-31)). Studies are required to determine if the replacement with a different lexeme is simultaneous or consecutive. Semantic knowledge is used in the compression of linguistic unit into a more concise form (see example (8-32)).

(8-29) In contests staged between two first-instar larvae, … [R-2-1]

→ When fighting takes place between two first instars, …

[A-1-5; bes1-9639061]

(8-30) the female was secured in the male's grappling legs, … [R-1-5]

→ the wriggling female is restrained in the male's grasp. [A-1-3: bes1-9946164]

(8-31) shrubland vegetation in cismontane southern California [I-2-1]

→ habitat in southern California; [A-1-1; oec1-99120304]

(8-32) trees were sampled [five times at 1-month intervals on 28 June, 30 July, 27 August, 24 September and 29 October 1994]$_{NP}$. [M-4-6]

→ This study tested the effect of foraging by ants, … on spider assemblages in Douglas-fir canopies in a [5-month]$_{A}$ ant-exclusion experiment. [A-1-3; oec2-97109313]

8.5 Discussion and Concluding Remarks

An alternative way to condense is to compress complex linguistic units into simpler ones. Compression may be divided into two main groups: (a) by deletion of support verb, catenative from verbal complexes and verb from prepositional verb, and (b) by derivation and compound noun formation. In both cases, lexical units are deleted. The latter is more commonly encountered and effective at reducing text. While the former is not commonly encountered, only one example per two documents, it still merits an investigation as it appears to involve determinate situations, and may be pertinent in other types of corpus. While catenatives are often associated with phasal verbs, e.g. to *start/continue/stop to VERB*, only one example was found in the study. The examples in our corpus involve lexical verbs of a particular semantic category that expresses uncertainty, e.g. *tend to*, *appear to*, *seem to*, etc. Mmeanwhile, we observed the compression of

While the above phenomena have been observed in linguistics, we found the compression of prepositional verbs particularly interesting. Studies are required for all the compression processes to determine the context or the list of catenatives that may be safely deleted without problems.

We provide here a good example of expansion and compression from our corpus. *Parasitoids* is expanded into *presence of parasitoids*, and *appeared to be more important* and *exploitation of food resources by Argentine ants* respectively compressed into *explained … far better* and *ant foraging*.

(8-33) Overall, parasitoids$_w$ appeared to be more important$_x$ than temperature in inhibiting the exploitation of food resources by Argentine ants$_z$ in Brazil. [I-4-4]

→ Overall, the presence of parasitoids$_w$ explained$_x$ observed variation in Argentine ant foraging$_z$ far better$_x$ than temperature, [A-1-10; oec2-98117420]

Chapter 9

Aggregation with and without Explicit Signs

Studies on sentences combining or aggregation often involve the use of connectives and short made-up sentences. While there are notable contributions on aggregation from text generation (see Dalianis, 1999; Shaw, 1998; Dalianis & Hovy, 1993), the work is of little immediate benefit to summarization by sentence extraction. The simple made-up sentences do not reflect the complex state of affairs in documents such as scientific and technical journal articles with urgent need for abstracts. The same holds true for potential contributions in grammar books (see Cattell, 1969; Leech & Svatvik, 1975); the sentences combined too do not come close in complexity with those actually written (see Table 9-1 below).

Table 9-1. Sentences Aggregated in Studies from Text Generation and from Grammar Books

Examples from Text Generation	Examples from Grammar Books
Mary sold tomatoes on Monday. Mary purchased cars on Tuesday. ... Mary had a garage sale on Sunday. John sold tomatoes on Monday. John purchased cars on Tuesday. ... John had a garage sale on Sunday. → Mary and John each did business all week. [from Dalianis (1999:386)]	He heard an explosion. He phoned the police. → He heard an explosion and (he) phoned the police. OR → When he heard an explosion, he phoned the police. [from Leech & Svartvik (1975:288)]

If studies on summarization are to benefit real applications, then research must reflect real contexts. For a recent work on producing concise sentences, see Jing & McKeown (2000) and Saggion (2000). In the aforementioned studies, sentences are almost always aggregated with the use of an explicit sign, a connective or a (semi-)colon. But explicit signs restrict the number of units that may be combined at any one time. So, how does information during abstracting get condensed into fewer units without excessive use of connectives?

From a comparison of sentences in document selected for abstracting and the abstract, this reconnaissance study on entomology-related articles provides some data on aggregation in section 9.1, and in section 9.2, reports on some preferred patterns in aggregation of authors when writing abstracts for their journal articles. More sentences were aggregated without than with the use of an explicit sign, such as a connective or a (semi-)colon. The chapter also discusses some prerequisites and difficulties anticipated for an abstracting system.

9.1 Some Data on Aggregation

9.1.1 Distribution

About 37% of ab-sentences in the study corpus were aggregated from two ft-sentences, while about 27% were constituted from three or more sentences.

Table 9-1. Distribution of ft-sentences to construct an ab-sentence

Sub-corpus (no. ab-sn.)	No. of ft-sentence (%)			
	1	2	3	≥ 4
bes1 (120)	43 (35.83)	48 (41.67)	20 (16.67)	9 (7.50)
bes2 (120)	43 (35.83)	48 (40.00)	14 (11.67)	15 (12.50)
oec1 (136)	43 (31.62)	45 (33.09)	30 (22.06)	16 (11.77)
oec2 (158)	60 (37.97)	56 (35.44)	25 (24.05)	14 (8.86)
Corpus (534[+])	189 (35.39)	197 (36.89)	89 (16.67)	54 (10.11)

[+] Five ab-sentences did not have matches.

9.1.2 Source of Sentences in Document

To simplify the study, we only looked at the simplest case of two-sentence aggregation. Most sentences aggregated were from the same section with Introduction as the highest contributor, and Method, the lowest.

Our study also revealed that sentences involved in two-ft-one-ab matches were more likely to be from the same section (see Table 9-2). When from different sections, the sentences were likely to be from Results and Discussion.

Table 9-2. Distribution of selected ft-sentences in two-ft-one-ab-sentence construction

Section	Introduction	Method	Results	Discussion
Introduction	57[+](28.9)			
Method	12 (6.1)	10 (5.1)		
Results	5 (2.5)	9 (4.6)	31 (15.7)	
Discussion	11 (5.6)	5 (2.5)	27 (13.7)	30 (15.2)

[+] No. of sentences (percentage)

Eighteen percent of ab-sentences has its source in sentences that were immediately adjacent. The implication of this finding for aggregation is that adjacent sentences are more likely to be on the same topic than sentences from different paragraphs/sections, and the anaphor is more likely to refer to an element mentioned in the preceding sentence.

9.2 Categorization of aggregation

Reape & Mellish (1999:23-25) proposed a four-category typology. Conceptual aggregation was distinguished from semantic and lexical aggregations. While the latter two presumptively involve linguistic knowledge, the examples given do not appear to be far different from that of conceptual aggregation which implicates world/domain. However, on the basis of whether an explicit sign was used or not, we propose three categories of aggregation. If the explicit sign is a connective or (semi-)colon, then CONNECTIVE or (SEMI-)COLON respectively, and if no sign was used, then CONFLATION. In the last category of CONFLATION, the basis of aggregation is knowledge, linguistic or world/domain. See Table 9-3 to see how our proposed categorization compares with that by Reape & Mellish (*ibid.*). Each of these categories, C1-C3, is discussed below.

Table 9-3. Categorizations proposed by present study vs. Typology of aggregation surveyed by Reape & Mellish (1999)

Proposed category	Reape & Mellish's typology
By conflation	Conceptual aggregation, Semantic aggregation, Lexical aggregation, Referential aggregation,
With connective	Discourse aggregation, Syntactic aggregation,
With (semi-)colon	-

9.2.1 By Conflation

Seventy-five percent of two-sentence aggregations were the result of *conflation*[60] (see **text in bold**). Two semantically equivalent text units may be conflated by: (a) splicing and joining, or (b) merging them. Units are merged on the basis of semantic similarity. Often one sentence (Sx) is used as the main sentence.

C1a: $[X_1Y]_{Sx} + [X_2Z]_{Sy} \rightarrow [X_2Y]_S$	$\mid$ 'X_1' ≅ 'X_2'[61];

In example (9-1), text unit ***small, early-instar bolas spiders*** was spliced off one sentence and joined to text unit *of both sexes attract moth flies in the genus Psychoda* in main sentence [I-3-6].

(9-1) **Small, early-instar bolas spiders** ~~do not capture moths~~. [I-3-1]

juvenile bolas spiders of both sexes attract adult male flies in the genus *Psychoda*. [I-3-6]

→ **Small, early-instar bolas spiders** of both sexes attract moth flies in the genus *Psychoda*, ... [A-1-5; oec1-97112572]

C1b: $[X_1Y]_{Sx} + [X_2Z]_{Sy} \rightarrow [XY]_S$	$\mid$ 'X_1' + 'X_2' ≅ 'X'

[60] To conflate = "to combine two or more things to form a single new thing" (LDOCE, 1995).
[61] X, Y, Z are units of text, and 'X' = meaning of X.

In example (9-2), sentences are aggregated when semantically equivalent text units were merged, before being optionally followed by other condensation sub-processes, such as deletion (~~deleted text~~) and substitution. Units are aggregated without any explicit use of a connective, or a (semi-)colon.

recent study + field studies

→ *recent field studies*

→ *preliminary field observations*

(9-2) ~~A~~ **recent study** ~~of the life history of this annual species~~ revealed an ~~unusually~~ extended reproductive period, which results in a very wide ~~and possibly bimodal~~ size distribution of ~~the coexisting~~ juvenile instars. [I-6-2]

Field studies ~~have suggested that size difference might be important in wolf spider cannibalism~~. [D-1-4]

→ **Preliminary field observations** indicated an extended reproductive period, which results in a very wide size distribution of juvenile instars. [A-1-3; bes1-9945349]

Aggregations, however, are rarely as direct as examples (9-1) and (9-2). In example (9-3), anaphor resolution is required: ***species*** is the lexical anaphor for ***ants and spiders***.

(9-3) **Ants and spiders** are ~~among the most ubiquitous and diverse predators in terrestrial ecosystems~~. [I-1-1]

Many species ~~share the same trophic level and~~ can potentially compete with and prey upon each other. [I-1-2]

→ **Spiders and ants** are potential competitors and mutual predators. [A-1-1; oec2-97109313]

In example (9-4), experimental knowledge is first required to know that text unit ***CO_2 sensitivity*** is a metonym for text unit ***sensory organs that are specialised to the detection of CO_2***, before a unit was selected. The selected unit was transformed finally to ***sensory organs that detect CO_2*** in the abstract.

(9-4) ~~Surprisingly, however~~, **sensory organs that are specialised to the detection of CO_2** find their strongest expression in ~~the almost exclusively~~ herbivorous Lepidoptera. [I-1-7]

~~This suggests that~~ **CO_2 sensitivity** is important throughout that order, but the functional role has remained unclear. [I-1-8]

→ **Sensory organs that detect CO_2** are common in herbivorous moths and butterflies, but their function has been unclear **until now**. [A-1-1; oec2- 97110539]

Hyponym(lepidoptera) = moth, butterfly (WN)
Syn(role) = function (WN)

Note the simultaneous occurrence of other condensation processes, namely substitution with a less technical term: *moths* and *butterflies* → *Lepidoptera*, and compression into fewer words: *functional role* → *function* (see example (9-5)).

(9-5) ~~If insect species without information about their host range are excluded,~~ **most of the recorded species are oligophagous**, feeding on more than two genera within the Brassicaceae. [R-2-1]

Irrespective of the feeding niche, **oligophagous species** dominate the insect fauna in the Brassicaceae, ~~whereas specialized species dominate the fauna of the Cardueae~~. [D-2-6]

→ Irrespective of the feeding niche, **species are predominantly oligophagous**, feeding on more than two Brassicaceae genera. [A-1-4; oec2-98113391]

9.2.2 With a Connective

Leech & Svartvik (1975:158) listed: coordination, subordination, and adverbial link, as three ways to aggregate clauses. Depending on whether equal, or unequal weight is to be given to the units, the appropriate conjunction, or adverbial is then used.

9.2.2.1 By Coordination

The most common way to aggregate, is with a coordinate conjunction, e.g. *and*, *but*, *or*.

C2a: $[S_1]_S + [S_2]_S \rightarrow [S_1 \text{ connective } S_2]_{Sc}$

Selected clauses from complex sentences (Sc) are joined. It is not necessary that there be a shared unit.

(9-6) ~~Facultative slavemakers are able to forage, nurse their brood and construct their nest like free-living ants, and hence~~ colonies without slaves are common. [I-1-3]

Formica subnuda is a facultative slave-making ant, ~~and belongs to the *F. sanguinea* group~~. [I-2-1]

→ *Formica subnuda* is a facultative slave-making ant**, and** colonies without slaves are often found. [A-1-1; bes2-9638145]

(9-7) Its invasions ~~threaten endemic arthropods in Hawaii and~~ eliminate native ants ~~in California, Australia, and South Africa~~. [I-1-4]

Argentine ants also tend homopterans ~~and augment their destructiveness in agriculture~~. [I-1-6]

→ In its introduced ranges it eliminates native ants **and** tends agricultural pests. [A-1-2; oec2-98117420]

As sentences studied are highly complex with multiple sentences, it is possible that the units aggregated are from the same sentence.

(9-8) ~~The four … stimuli derived from this video had different degrees of asymmetry, and were created to address different aspects of~~ asymmetry manipulation: ~~(1) removed: one tuft was removed~~, representing the most extreme level of FA or RA; ~~(2) reduced: one tuft was reduced in height such that the overall area was decreased by 25%~~, representing a mid-point within the range of natural FA variation; ~~(3) enlarged: …; (4) balanced: …~~ [M-6-3]

→ Asymmetry treatments represented values within the range of natural FA variation **as well as** more extreme values characteristic of regenerative asymmetry.

[A-1-9; bes1-9945087]

C2b: $[NP_1 VP_1]_{Sc} + [NP_1 VP_2]_{Sc} \rightarrow [NP_1\ VP_1$ connective $VP_2]_{Sc}$

If coordinated aggregation involves a shared unit, then the redundant unit has to be deleted. As in aggregation by conflation, to combine, the abstractor must first determine the units to be equivalent or synonymous: in example (9-9), *parasitism by eulophids* and *eulophid parasitism* are equivalent.

(9-9) ~~*Phyllonorycter* survival~~, parasitism by eulophids, ~~and unknown causes of mortality~~ varied significantly among ~~naturally occurring~~ hybrid and parental plants in 1994. [D-1-2]

Eulophid parasitism, ~~rather than unknown mortality~~, appeared to account for the variation in survival among taxa. [D-1-3]

→ Parasitism by eulophid wasps differed significantly among taxa in 1994 **and** appeared to account for the variation in their survival. [A-1-3; oec2-97110360]

Aggregation was followed by a substitution which requires domain knowledge: generic word *taxa* substitutes for *hybrid and parental plants*.

In example (9-10), aggregation is complicated by: anaphor resolution; and knowing when and what may be deleted.

(9-10) Simple movement rules, ~~such as the two rules described above~~, may be acquired through a gradual associative learning process, ~~such as the learning mechanisms which lead to the formation of flower-species preferences~~. [I-3-1]

An alternative hypothesis is that these are innate, ~~instinctive~~ processes, ~~and thus should be observable in bees with no previous foraging experience~~. [I-3-2]

→ These patterns may be innate, **or** they may be learned through the bees' early foraging experience. [A-1-2; bes2-9639381]

While the fact that the sentences here are consecutive, helps to determine the entity referred to by the anaphor *these*, document knowledge is still required to determine what the noun referred to is. Is it *rules*, or is it *process*?

9.2.2.2 By Subordination

The patterns of aggregation for subordinated and coordinated aggregation differ in the choice of conjunction which depends very much on the communicative intent of the author which a non-author abstractor usually has no direct access.

(9-11) ~~Combined, these two findings suggest that~~ *S.dumicola* has control over its mean sex ratio but not of its variance. [D-5-4]

~~There are two possibilities that are not mutually exclusive: either~~ the sex ratio biasing mechanism in *S.dumicola* cannot be modified to control the sex ~~of individual offspring~~ or the sex ratio variance ~~is selectively neutral in this system~~. [D-6-2]

→ The sex ratio biasing mechanism in this species, **therefore**, apparently only allows control of the mean sex ratio but not of its variance. [A-1-7; bes1-9946237]

9.2.3 With a Semi-Colon or a Colon

In aggregations with a semi-colon or colon, the punctuation is substitutes for the implicit semantic relation which has been expressly omitted. Aggregations with a (semi-)colon as with other aggregation types, are accompanied by various condensation sub-processes.

9.2.3.1 With a Semi-colon

Ehrlich & Murphy (1974:111) say that "When no close relationship exists between two independent clauses, a semicolon can be used to join them".

C3: $[X]_S + [Y]_S \rightarrow [X$ (semi-)colon $Y]_S$

While this makes the semicolon a convenient means for combining just about any two clauses, most of the clauses aggregated in the present study are related.

(9-12) The most abundant prey organisms brought to the nest were Aphidoidea (48.1%), followed by Psocoptera (12.5%), ~~and Lepidoptera larvae (6.0%)~~. [R-9-4]
Only three spiders (two lycosids and one salticid) were brought to the nests. [R-9-5]
→ The majority of prey captured by ants were Aphidoidea (48.1%) and Psocoptera (12.5%)**<semi-colon>** spiders represented only 1.4% of the ants' diet.
[A-1-8; oec2-97109313]

In one of four cases, the (semi-)colon is additionally accompanied by a connective to make explicit the semantic relation (see examples (9-13) and (9-14)).

(9-13) ~~*E.snoddyi* males that obtained copulations and unsuccessful males that did not obtain copulations were analyzed to determine if~~ male body size or male balloon size were important criteria for male mating success. [R-4-1]
The empty balloon produced by some species of empidine flies has been hypothesized to be a sexually selected trait. [D-4-1]
→ Both male body size and balloon size are important components in determining male mating success; **<semi-colon> however,** the empty balloon does not appear to play a typical role as a sexually selected ornament. [A-1-11; bes1-9945161]

(9-14) This difference was not statistically significant, ~~and provided the basis for conclusions regarding~~ selective association among clone-mates during gall initiation. [R-2-4]

~~The data presented above indicate that~~ aphid foundresses do not discriminate actively between kin and non-kin during gall formation. [D-1-1]

→ There were no significant differences in the frequencies of communal gall occupation **<semi-colon> therefore**, active kin discrimination by *T.coweni* foundresses apparently does not play a role in their communal behavior, within the context of this experiment. [A-1-5; bes1-9843095]

9.2.3.2 With a Colon

Colons are used "to set off a series of words, phrases, or clauses from the rest of a sentence, to restate, explain or illustrate a statement immediately before it; ... to replace a semicolon for stylistic purposes [to break between clauses]" (Ehrlich & Murphy, 1974:25-27). In the study corpus, there were more examples of aggregation with a colon than with a semi-colon.

(9-15) ~~In this study, we investigate the role played by~~ a conspicuous male secondary sexual characteristic ~~in the courtship~~ of the wolf spider *Schizocosa ocreata* ~~(Hentz)~~ (Araneae: Lycosidae). [I-3-2]

~~Morphologically, these species can be distinguished only by~~ a male secondary sexual characteristic, a conspicuous tuft of bristles and dark pigmentation on the tibia ~~and patella~~ of the first pair of legs of mature male *S.ocreata*, ~~which is lacking in *S.rovneri* (as well as in the females and juveniles of both species)~~. [I-5-3]

→ Males of the brush-legged wolf spider, *Schizocosa ocreata* (Araneae: Lycosidae), possess a conspicuous male secondary sexual character**<colon>** dark pigmentation and tufts of bristles on the tibiae of their forelegs. [A-1-1; bes1-9638017]

9.3 Discussion

9.3.1 Occurrence of Aggregation

Aggregation is an important sub-process in condensation. Two-thirds of ab-sentences are the result of combining text units from different sentences.

Of two-sentence aggregations, three quarters were combined without an explicit sign by conflating semantically equivalent units, while the rest were combined with an explicit sign, a connective or a (semi-)colon (in the ratio of 4 :1). Most of the sentences aggregated come from Introduction, which is reflective of the section where important sentences might be found.

9.3.2 Types of aggregation

9.3.2.1 By Conflation

While Shaw (1998:139) in his study on text generation noted "coordinate constructions [to be] the most popular aggregation operations, followed by PPs, and then adjectives", three per four ab-sentences in the present study were aggregated by conflation. This is not surprising since aggregating with an explicit sign (connective/(semi-)colon), restricts the number of units that may be combined at any one time. For maximum condensation of information into a single unit, aggregation by conflation is more effective.

To conflate, a myriad of processes, condensation and non-condensation, is implicated, and multiple sentences are often involved. One quarter of ab-sentences were aggregated from three or more sentences. Also, the units must first be determined to be semantically equivalent. As the units are often equivalent under the guise of synonyms, hypernyms, partial repetitions and metonyms, knowledge ranging from linguistic to experimental to world/domain, is prerequisite.

9.3.2.2 With Connective or (Semi-)Colon

To help a reader process a complex sentence on unfamiliar material, aggregations with connectives which make explicit the semantic relation between units joined, are preferred. In such aggregations, even if a non-author abstractor can decide on the pattern of aggregation, the crux of the problem is which conjunction or adverbial to use such that author's intent is communicated.

The use of a (semi-)colon to aggregate sentences does not mean that there is no semantic relation between the sentences, rather, that "the connection is implicit, and has to be inferred by the reader" (Leech & Svartvik, 1975:162). The non-explicit mention of a semantic relation could also be author's way of compelling a reader to participate in the development of the text. However, because of the need to be explicit in scientific and technical texts, an adverbial or conjunction is additionally inserted 25% of the time to help a reader process unfamiliar text.

Unlike linguistic units combined in made-up sentences, units actually aggregated are not only different in syntactic class, but from sentences of differing structure and require experimental knowledge to know that they co-refer. Compare made-up sentences in aggregation in Fig. A7-1 in Appendix VII, and sentences actually written (see examples (9-16) and (9-17)).

(9-16) Plant hybridization affects tritrophic-level interactions ~~in this system in the field~~. [D-1-1]

~~However~~, the common garden results ~~strongly~~ suggest that the differences in enemy impact among plants has a genetic basis. [D-6-3]

→ The common garden results show that genetic differences in plants affect the herbivore-parasitoid interaction. [A-1-7; oec2-97110360]

(9-17) ~~As foragers~~, I used penultimate-instar female crab spiders *Misumena vatia* (Thomisidae) ~~collected within the preceding few hours~~ from flowers of two species frequented by these spiders: ox-eye daisy *Chrysanthemum leucanthemum* and common buttercup *Ranunculus acris*. [I-4-1]

M.vatia are sit-and-wait predators that hunt primarily on flowers. [I-4-2]

→ *M.vatia* is a sit-and-wait predator, and the two flower species used, ox-eye daisy *Chrysanthemum leucanthemum* and common buttercup *Ranunculus acris*, are important hunting sites. [A-1-3; oec1-99120252]

9.3.3 Problems and Prerequisites

Because sentences in scientific and technical documents are not only long[62], but complex, a simple aggregation is not possible. To ensure that the output sentence is readable, aggregation is almost always accompanied by various condensation sub-processes, e.g. deletion, to prune off marginal texts. Note that the abstract has to be formulated in words appropriate to the readership.

Besides these problems, an abstracting system is faced with problems related to the prerequisites of abstracting, such as anaphor resolution, determination of the entity referred to in metonymy, and determination of the full form of partial repetitions. Even if the problem of anaphor resolution is alleviated when the sentences are adjacent, and the full form of compound nouns can be determined by a simple concordance of relevant nominal forms, the uncovering of an entity referred to in metonymy which requires experimental or world knowledge, remains problematic. To go beyond, solutions to these problems have first to be found.

9.4 Concluding Remarks

As just seen, aggregation in real situations is far different from that treated in hypothetical situations. While one may know how to aggregate, and to detect redundancy, the role of experimental and domain knowledge in conflation is equally urgent. Because conflation is an effective and common means of aggregation, future studies should look into the exploitation of knowledge to this end. Pending long-term measures to understand this condensation sub-process, short-term studies can concentrate on condensing single sentences with the ultimate aim of combining them.

A proposed study situation is scientific and technical articles, which not only have a high turnover and demand, but are a source of examples for finding strategies/patterns in aggregation. As aggregation involves other condensation sub-process, parallel studies should be conducted to address problems on the (automatic) deletion, and identification of synonymous units, and the entity referred to in metonymy.

---

[62] With an average of 22 words per sentence.

Chapter 10

Conclusion and Future Work

In our survey of research on automatic summarization, the question that struck us most was why we had not been able to proceed beyond content selection. A closer examination of the research showed the reason to be because of the great attention paid to the product and little to the processes leading to its production. While there is some work on improving the readability of extracted abstracts, the focus still remains on the product.

To clear up the fuzzy comprehension of condensation, we first separated out, on the basis of the operation involved, the assortment of terms and processes proposed by various researchers in summarization and text generation. To this initial categorization, we next augmented other processes that the English language has for expressing content concisely, and also processes identified from our comparative study of sentences from full text used by an author in abstracting and the abstract. The result of this exercise is a provisional four-category typology of condensation sub-processes: generalization, deletion, compression and aggregation, given earlier in Chapter 5. Definitions from linguistics are proposed for the condensation sub-processes identified. The aim is for the definitions to serve as guide to distinguish between sub-processes which form the basis of condensation. Future work is needed to investigate how these basic processes interact with which process(es) and under what conditions.

In our study, we looked into the linguistic units involved for each process. A range of units in replacement was identified. Author-abstractors were found to reuse the stems of words in full text during abstracting. About 55% of stems in abstract are found in full text. In the scientific corpus investigated, authors showed a tendency to replace technical words with general words. As there are only so many general words in the English language which may substitute for technical words, biology has its own sublanguage device to create domain synonyms by postposing a generic word to a derived form of a technical word, e.g. *kleptobiont* → *kleptobiotic spider*. When abstracting documents from particular domains, it helps to know domain-related devices for creating replacements.

Replacement with synonyms while not common, appears to be restricted to adjectives and adverbs, which play a marginal role in the text. Few substitutions with synonym involve nouns; most substitutes for nouns are retained in same stems. WordNet appears to be a reliable resource for finding synonyms even those involving technical words. However, a problem that remains is which among the possible senses to use. In the scientific corpus investigated, numerical expressions were consistently rendered less precise; absolute numbers were rounded off, and fractions and ratios used as replacements for percentages. While less precise, the reformulation in content is no less accurate in a context where only the gist of a document's content need be re-conveyed.

Deletion is an obvious way to condense text. Linguistic units commonly deleted include illocution markers containing an author's overt presence first person pronouns. Connectives which lose their function when sentences are extracted from their context are almost always deleted. Parenthetical texts, apposed texts and repetitions are also commonly deleted. While deletion of such linguistic units may be a small first step in condensation, multiple deletions of such units alone can significantly abridge a text without critical loss in core content. Because parenthetical and apposed texts are set off by punctuation, and first person pronouns *I* and we, connectives and repetitions are easily recognizable, these linguistic units are most attractive as considerations to computational linguists to operationalize for inclusion in abstracting systems. However, note that while the overt presence of author may be omitted, some degree of first person pronouns are for various reasons retained.

Among linguistic units deleted during abstracting are those that are emphatic, or implicit. Deletion of these units are a challenge to future work as various kinds of knowledge domain, world and experimental are involved. Equally instructive are studies into the identification of dispensable units, and the extent to an author's overt and non-overt presence may be deleted. Pending such long-term studies, short-term projects may concentrate on finding lists of linguistic units that may be deleted to complement summarization by sentence extraction.

Two linguistic units in the English language that are compressible to fewer essential units are verbal complexes containing a support verb, or a catenative. While these units do not occur with great frequency in the particular scientific corpus investigated, they appear to involve

determinate situations, and are therefore operationalizable. Comprehensive studies with other types of corpus are required to understand this observation, or if it may be generalized for texts of a scientific and technical nature.

As seen from our study, aggregation in real situations is far more complex from that treated in hypothetical situations. While aggregation is an important condensation process: two-thirds of sentences in abstract has its source information in multiple sentences, connectives was not the main means by which sentences are combined. In the simplest of aggregation involving just two sentences, three-quarters of sentences were the result of conflation, i.e. merging of semantically equivalent units. Because conflation is an effective and common means of aggregating segments of text without the use of excessive connectives or colons or semi-colons, studies into the exploitation of knowledge are needed to this end.

From our restricted study in a scientific domain and whatever available information on concise reformulation of content, we have drawn up a typology of its sub-processes for consorted use of terms and identified some transformations that may be operationalized. While our study with author-written abstracts is plausibly more elaborate than that which may be applied by a system, we are interested in the condensation devices. Our study which is restricted to just one type of scientific documents needs to be extended to other document types to determine their domain-related devices for condensation.

When abstracting in a restricted domain, some substitutions involve synonymous technical wordforms. As such unlikely synonyms are not to be found in any non-technical lexical resource, special resources need to be compiled from empirical studies. While we know that about one-fifth of the substitutes used during abstracting are hypernyms or words in varying lexical relations with a given word in full text, investigations are still needed to determine the contexts and factors affecting generalization. To know when to apply substitution with same stem and when to apply substitution with synonyms, an investigation into the correlation between substitution with same stem for important content and substitution with synonyms for marginal content is needed. These studies should parallel other studies on summarization.

Because of the tedium of carrying out matches, a semi-automatic matching procedure is needed to help pick out candidate sentences to encourage research on this condensation process of abstracting. For a full appreciation of the complexity of the problem, we encourage studies preferably on abstracts prepared by an author himself.

In our study of abstracting, we saw that linguistic units were not only compressed and deleted, but also expanded or inserted or apposed to compact more information into the abstract. The latter calls for a review of the general meaning of condensation, the re-expression of content in fewer words. With a better comprehension of the sub-processes in content condensation, the next stage in summarization research would be on the interplay the processes. We attached in Appendix VIII, four documents with varying lengths of abstract (the longest, one intermediate, and two shortest) to illustrate the complexity of the problem. Document VIII-1 is the longest abstract of 21 sentences in the document. Document VIII-3 and VIII-4 have the shortest abstract of five sentences each in the document. Document VIII-2 is of intermediate length.

From the full text sentences identified to have been used in abstracting, we were able raise doubts on the unreliability of a couple of cues and confirm the reliability of "paragraph feature" commonly used in sentence selection. The latter implies the exploitation of text structure in abstracting. However, this observation should be verified with other corpus type. While we prefer a more linguistic-based approach to sentence selection, we propose the inclusion of distribution of important sentences over sections as a feature to current statistical techniques for abstracting structured documents. The proportion of sentences to select over section may however need to be verified for different corpus type. Meanwhile because of the ease with which cues or fixed phrases may be exploited, we propose that these indicators be used to choose between the more likely of candidate sentences for selected in abstracting via a linguistic-based process, although not as basis of sentence selection.

Bibliography

Alley, M. (1996) The Craft of Scientific Writing. New York: Springer-Verlag. 3rd edn.

American National Standards Committee Z39. (1979) Terms Defined in Z39: Published and Draft Standards. Washington, D.C.: American National Standards Institute.

Barzilay, R. & Elhadad, M. (1997) Using Lexical Chains for Text Summarization. Proc. of the ACL '97/EACL'97 Workshop on Intelligent Scalable Text Summarization, pp 10-17, Universidad Nacional de Educación a Distancia, Madrid, Spain, July 11.

Baxendale, P.B. (1958) Man-made Index for Technical Literature - An Experiment. IBM Journal of Research and Development, 2(4):354-361.

Benbrahim, M. & Ahmad, K. (1995) Text Summarisation: The Role of Lexical Cohesion Analysis. The New Review of Document and Text Management. 1:321-335.

Brandow, R., Mintze, K. & Rau, L.F. (1995) Automatic Condensation of Electronic Publications by Sentence Selection. Information Processing and Text Management. 31(5)675-685.

Bussman, H. (1996) Dictonary of Language and Linguistics. London: Routledge.

Cattell, N.R. (1969) The New English Grammar. Cambridge, MA: The MIT Press.

Crystal, D. (1997) Dictionary of Linguistics and Phonetics. 4th edn. Oxford: Blackwell.

Dalianis, H. & Hovy, E. (1993) Aggregation in Natural Language Generation. In G. Adorni and M. Zock (eds.) Trends in Natural Language Generation: An Artificial Intelligence Perspective. Fourth European Workshop, WENLG '93, Pisa, Italy, April 1993. pp 88-105.

Dalianis, H. (1999) Aggregation in Natural Language Generation. Computational Intelligence. 15(4):384-414.

Davis, E.B. & Schmidt, D.(1998) The Biological Literature. In A. Kent & H. Lancour (eds.) Encyclopedia of Library and Information Science. New York: Marcel Dekker Inc.

Edmundson, H.P. (1969) New Methods in Automatic Abstracting. Journal of the ACM, 16(2):264-285.

Ehrlich, E. & Murphy, D. (1974) Concise Index to English. New York: McGraw-Hill Book Company.

Fellbaum, C. (1998) A Semantic Network in English Verbs. In C. Fellbaum (ed.) WORDNET: An Electronic Lexical Database. Cambridge, MA: The MIT Press. pp 69-99.

Fellbaum, C. (ed.) WORDNET: An Electronic Lexical Database. Cambridge, MA: The MIT Press.

Halliday, M.A.K. & Hasan, R. (1976) Cohesion in English, London, UK: Longman.

Halliday, M.A.K. (1993) Some Grammatical Problems in Scientific English. In M.A.K. Halliday & J.R. Martin Writing Science: Literacy and Discursive Power. Pittsburg, PA: University of Pittsburg Press. pp 69-85. [originally published in Australian Review of Applied Linguistics: Genre and Systemic Functional Studies, (1989), 5(6):13-37.]

Hirschman, L. & Sager, N. (1982) Automatic Information Formatting of a Medical Sublanguage. In R. Kittredge & J. Lehrberger. Sublanguage: Studies of Language in Restricted Semantic Domains. Berlin/New York: Walter de Gruyter. pp 27-80.

Hirst, G. & St-Onge, D. (1998) Lexical Chains as Representations of Context for the Detection and Correction of Malapropisms. In Christiane Fellbaum (ed.) WORDNET: An Electronic Lexical Database. Cambridge, MA: The MIT Press. pp 305-329

Hovy, E.H. (1990) Parsimonious and Profligate Approaches to the Question of Discourse Structure Relations. Proc. of the International Workshop on Natural Language Processing, pp 128-136. Dawson, Pensylvania.

Huddleston, R. (1988) English Grammar: An Outline, Cambridge: Cambridge University Press.

Jing, H. & McKeown, K.R. (2000) Cut and Paste Based Text Summarization. Proc. of the 1st Meeting of the North American Chapter of the Assoc. for Computational Linguistics, pp 178-85, Seattle, Washington, USA. April 29-May 4, 2000.

Kent, A. & Lancour, H. (1968) Encyclopedia of Library and Information Science. New York: Marcel Dekker, Inc. Vol. 1.

Kittredge, K. & J. Lehrberger (eds.) (1982) Sublanguage: Studies in Language in Restricted Semantic Domains. Berlin; New York: Walter de Gruyter.

Vande Kopple, W. (1985) Some Exploratory Discourse on Metadiscourse. College Composition and Communication, 36(1):82-93.

Vande Kopple, W. (1994) Some Characteristics and Functions of Grammatical Subjects in Scientific Discourse. Written Communication. 11(4):534-564.

Kupiec, J., Pedersen, J. & Chen, F. (1995) A Trainable Document Summarizer. Proc. of the 18th Annual International ACM SIGIR Conference on Research and Development on Information Retrieval, pp 68-73. Seattle, Washington. July 1995.

Leech, F. & Svartvik, J. (1975) A Communicative Grammar of English. London: Longman.

Longman Dictionary of Contemporary English. (1995) Essex, England: Longman Group Ltd. 3rd edn.

Luhn, H.P. (1968) The Automatic Creation of Literature Abstracts. In C.K. Schultz (ed.) H.P. Luhn: Pioneer of Information Science, Selected Works. New York: Spartan Books. pp 118-125. [originally published in IBM Journal of Research and Development, 1958, 2(2):159-165.]

Mani *et al.* (1998) The TIPSTER SUMMAC Text Summarization Evaluation. Mitre Technical Report. MTR 98W0000138. McLean, VA: The Mitre Corporation.

Mann, W.C. & Thompson, S.A. (1987) Rhetorical Structure Theory: A Theory of Text Organization. Report ISI/RS-87-190, Information Sciences Institute, University of Southern California. [Also, in Mann, W.C. & Thompson, S.A. (1988) Rhetorical Structure Theory: Toward a Functional Theory of Text Organization. Text 8(3):243-281.]

Marcu, D. (1997) From Discourse Structures to Text Summaries. Proc. of the ACL'97/EACL'97 Workshop on Intelligent Scalable Text Summarization, pp 82-88, Universidad Nacional de Educación a Distancia, Madrid, Spain. July 11.

Mathis, B., Rush, J.E., & Young, C.E. (1973) Improvement of Automatic Abstracts by the Use of Structural Analysis. Journal of the American Society for Information Science, pp 101-109.

Matthews, P.H. (1997) Oxford Concise Dictionary of Linguistics. Oxford: Oxford University Press.

Maybury, M.T. (1995) Generating Summaries from Event Data. Information Processing and Management, 31(5):735-751.

McArthur, T. (ed) (1992) The Oxford Companion of the English Language. Oxford: Oxford University Press.

Melcuk, I.A. (1996) Lexical Functions: A Tool for the Description of Lexical Relations in a Lexicon. In L. Warner (ed.) Lexical Functions in Lexicography and Natural Language Processing. Amsterdam/Philadelphia: John Benjamins Publishing Company. pp 37-101.

Melcuk, I.A., Clas, A. & Polguère, A. (1995) Introduction à la Lexicologie Explicative et Combinatoire. Belgium: Duculot.

Meyers, G. (1992) 'In this paper we report ...': Speech acts and scientific facts. Journal of Pragmatics. 17:295-313.

Meyers, G. (1996) Strategic Vagueness in Academic Writing. In E. Ventola & A. Mauranen (eds.) Academic Writing: Intercultural and Textual Issues. Amsterdam/Philadelphia: John Benjamins Publishing Company. pp 3-17.

Miller, G.A. (1998) Nouns in WordNet. In C. Fellbaum (ed.) WORDNET: An Electronic Lexical Database. Cambridge, MA: The MIT Press. pp 23-46.

Miller, K.J. (1998) Modifiers in WordNet. In C. Fellbaum (ed.) WORDNET: An Electronic Lexical Database. Cambridge, MA: The MIT Press. pp 47-68.

Morris, J. & Hirst, G. (1991) Lexical Cohesion Computed by Thesaural Relations as an Indicator of the Structure of Text. Computational Linguistics, 17(1):21-48.

Ono, K., Sumita, K. & Miike, S. (1994) Abstract Generation Based on Rhetorical Structure Extraction. Proc. of the International Conference on Computational Linguistics. (COLING-94), pp 344-348. Japan.

Paice, C.D. & Jones, P.A. (1993) The Identification of Important Concepts in Highly Structured Technical Papers. Proc. of the ACM-SIGIR '93, pp 69-78.

Paice, C.D. (1990) Constructing Literature Abstracts by Computer: Techniques and Prospects. Info. Processing & Management 26(1):171-186.

Polguère, A. (2000) Une base de données lexicales du français et ses applications possibles en didatique. Revue de Linguistique et de Didatique des Langues. 21:75-97.

Radev, D. & McKeown, K.R. (1998) Generating Natural Language Summaries from Multiple On-Line Sources. Computational Linguistics, 24(3):469-500.

Reape, M. & Mellish, C. (1999) Just What is Aggregation Anyway? Proc. of the 7th European Workshop on Natural Language Generation, pp 20-29, Toulouse, France, 13-14 May, 1999.

Rush, J.E., Salvador, R. & Zamora, A. (1971) The Abstracting and Indexing. II. Production of Indicative Abstracts by Application of Contextual Inference and Syntactic Coherence Criteria. Journal of the American Society for Information Science, pp 260-274.

Sager, N. (1982) Syntactic formatting of science information. In R. Kittredge & J. Lehrberger. Sublanguage: Studies of Language in Restricted Semantic Domains. Berlin/New York: Walter de Gruyter. pp 9-26.

Saggion, H. & Lapalme, G. (1998) Where does Information come from? Corpus Analysis for Automatic Abstracting. Rencontre Internationale sur l'éxtraction le Fltrage et le Résumé Automatique. pp 72-83, Sfax, Tunisie, 11-14 November 1998.

Saggion, H. (2000) Génération automatique de résumés par analyse sélective. PhD thesis. Université de Montréal.

Seuren, P. (1998) Automatic Summarization by Paragraph Initial Sentences Extraction. Rencontre Internationale sur l'extraction le filtrage et le Résumé Automatique, pp 64-71, Sfax, Tunisie, 11-14 November 1998.

Shaw, J. (1998) Clause Aggregation Using Linguistic Knowledge. Proc. of the 9th International Workshop on Natural Language Processing, pp 138-147, Niagara-on-Lake, Ontario, Canada, 5-7 Aug.

Sinclair, J. (1991) Corpus, Concordance, Collocation. Oxford: Oxford University Press.

Sparck Jones, K. & Endres-Niggemeyer, B. (1995) Automatic summarizing. Information Processing & Management, 31(5):625-630.

Sparck Jones, K. (1999) Automatic Summarising: Factors and Directions. In I. Mani and M. Maybury (eds.) Advances in Automatic Text Summarization, Cambridge, MA: MIT Press.

Task, R.L. (1997) A Student's Dictionary of Language and Linguistics. London: Arnold.

Upjohn, J., Blattes, S. & Jans, V. (1991) Minimum Competence in Scientific English, Grenoble: Presses Universitaires de Grenoble.

Appendix I

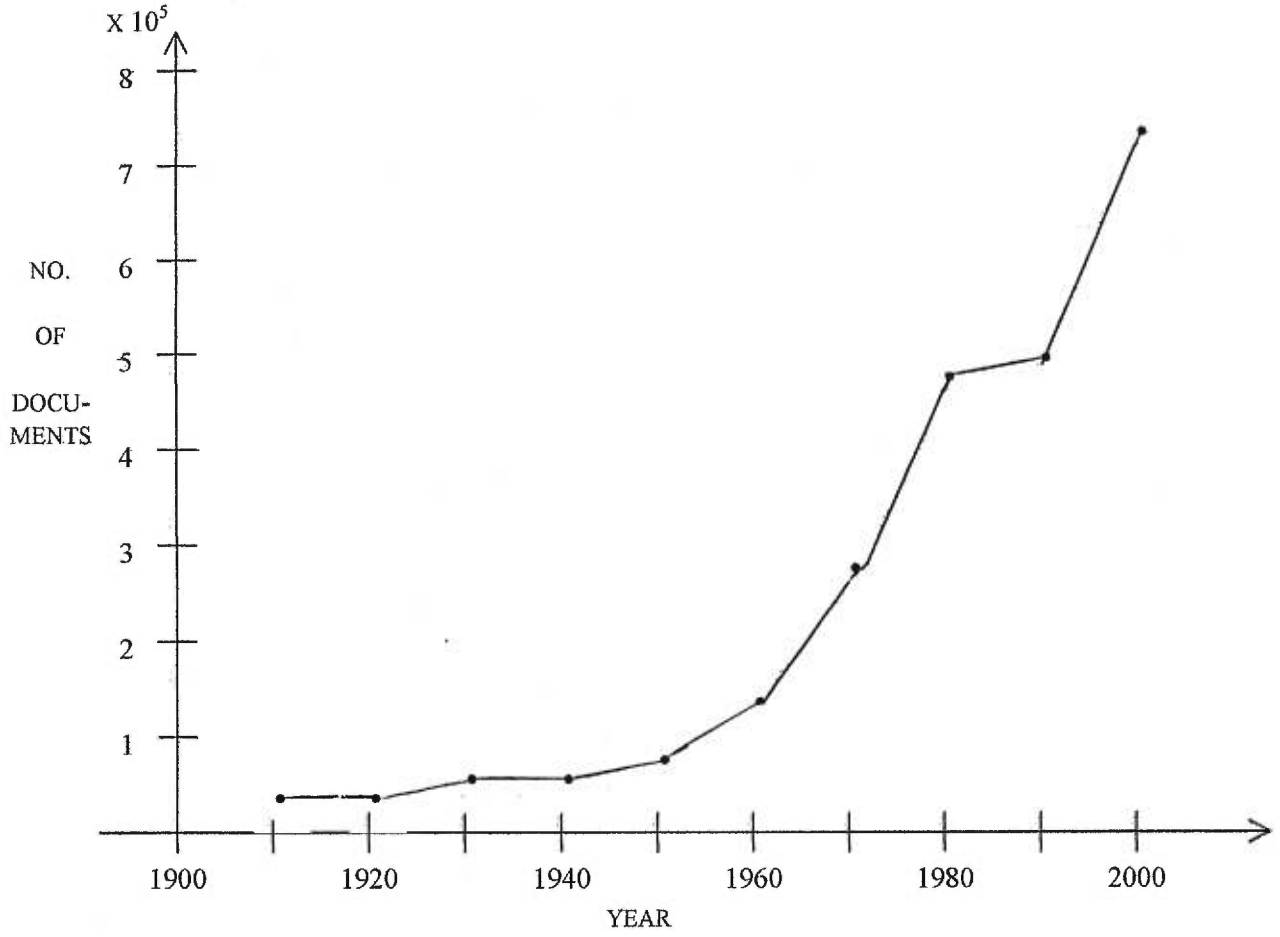


Fig A1-1. Number of documents abstracted by CAS over the period 1907-2000 (documents include patents, books, journal articles, dissertations, technical reports, and conference proceedings; see http://info.cas.org/EO/casstats.pdf for exact figures.)

Appendix II

Some definitions, namely 'abstract', 'abstracting' and 'summary', taken from Terms Defined in Z39, Published and Draft Standards. American National Standards Committee Z39. (1979) Washington, D.C.

ABSTRACT

(1) An abbreviated accurate representation of a document, without added interpretation or criticism and without distinction as to who wrote the abstract. Notes: A brief *review* of a document often takes on much of the character of an informative or informative-indicative abstract, but its writer is expected to include suitable criticism and interpretation. The word *synopsis* was formerly used to denote a resume prepared by the author as distinct from an *abstract* (condensation) prepared by some other person.

(2) A factual summary giving the significant content of a unit or publication (e.g. a scientific or scholarly paper, a technical report, a patent). It may accompany the full paper when originally published, or it may be issued separately with a citation referring to the original publication.

(3) An abbreviated accurate representation of the content of a work without added interpretation or criticism. The abstract should be accompanied by a bibliographic reference to the original work when reproduced separately from it.

(4) An abstract is usually much briefer than a synoptic and does not contain equations, tables, or figures to communicate results. Although good informative-type abstracts do summarize key findings, they seldom present sufficient specific information to permit direct application to those findings to another's work. Indeed, a synoptic contains an abstract.

ABSTRACTING

The practice of summarizing a scientific or scholarly paper or report in order to render in brief form the essential factual content.

SUMMARY

A summary is a restatement within a document (usually at the end) of its salient findings and conclusions, and is intended to complete the orientation of a reader who has studied the preceding text. Because other vital portions of the document (for example, purpose, methods) are not usually condensed into this summary, the term should not be used synonymously with "abstract"; that is, an abstract as defined above should not be called a summary.

Appendix III

Table A3-1a. List of Articles[+] in Sub-corpus bes1

bes1-9638017	Scheffer, S. J., Uetz, G.W. & Stratton, G.E. Sexual selection, male morphology, and the efficacy of courtship signalling in two wolf spiders (Araneae: Lycosidae). Behav Ecol Sociobiol (1996) 38:17–23.
bes1-9638253	Uetz, G.W., McClintock, W.J., Miller, D., Smith, E.I. & Cook, K.K. Limb regeneration and subsequent asymmetry in a male secondary sexual character influences sexual selection in wolf spiders. Behav Ecol Sociobiol (1996) 38:253–257.
bes1-9639061	Marris, G.C. & Casperd, J. The relationship between conspecific superparasitism and the outcome of in vitro contests staged between different larval instars of the solitary endoparasitoid *Venturia canescens*. Behav Ecol Sociobiol (1996) 39:61–69.
bes1-9740127	Fahey, B.F. & Elgar, M.A. Sexual cohabitation as mate-guarding in the leaf-curling spider *Phonognatha graeffei* Keyserling (Araneoidea, Araneae) Behav Ecol Sociobiol (1997) 40:127-133.
bes1-9842193	Pratt, S.C. Decentralized control of drone comb construction in honey bee colonies. Behav Ecol Sociobiol (1998) 42:193-205.
bes1-9843095	Miller III, D.G. Consequences of communal gall occupation and a test for kin discrimination in the aphid *Tamalia coweni* (Cockerell) (Homoptera: Aphididae). Behav Ecol Sociobiol (1998) 43:95-103.
bes1-9945087	Uetz, G.W. & Smith, E.I. Asymmetry in a visual signaling character and sexual selection in a wolf spider. Behav Ecol Sociobiol (1999) 45:87-93.
bes1-9945161	Sadowski, J.A., Moore, A.J. & Brodie III, E.D. The evolution of empty nuptial gifts in a dance fly, *Empis snoddyi* (Diptera: Empididae): bigger isn't always better. Behav Ecol Sociobiol (1999) 45:161-166.
bes1-9945349	Samu, F., Toft, S. & Kiss, B. Factors influencing cannibalism in the wolf spider *Pardosa agrestis* (Araneae, Lycosidae). Behav Ecol Sociobiol (1999) 45:349-354.
bes1-9946025	Saito, Y. & Sahara, K. Two clinal trends in male-male aggressiveness in a subsocial spider mite (*Schizotetranychus miscanthi*). Behav Ecol Sociobiol (1999) 46:25-29.
bes1-9946043	Heiling, A.M. Why do nocturnal orb-web spiders (Araneidae) search for light? Behav Ecol Sociobiol (1999) 46:43-49.
bes1-9946123	Kotiaho, J.S., Alatalo, R.V., Mappes, J. & Parri, S. Sexual signalling and viability in a wolf spider (*Hygrolycosa rubrofasciata*): measurements under laboratory and field conditions. Behav Ecol Sociobiol (1999) 46:123-128.
bes1-9946164	McLain, D.K. & Pratt, A.E. The cost of sexual coercion and heterospecific sexual harassment on the fecundity of a host-specific, seed-eating insect (*Neacoryphus bicrucis*). Behav Ecol Sociobiol (1999) 46:164-170.
bes1-9946237	Avilés, L., Varas, C. & Dyreson, E. Does the African social spider *Stegodyphus dumicola* control the sex of individual offspring? Behav Ecol Sociobiol (1999) 46:237-243.

[+] available at http://link.springer.de/link/service/journals/00265/bibs/

Appendix III (continued)

Table A3-1b. List of Articles[+] in Sub-corpus bes2

bes2-9638083	O'Donnell, S. RAPD markers suggest genotypic effects on forager specialization in a eusocial wasp. Behav Ecol Sociobiol (1996) 38:83–88.
bes2-9638145	Savolainen, R. & Deslippe, R.J. Slave addition increases sexual production of the facultative slave-making ant *Formica subnuda*. Behav Ecol Sociobiol (1996) 38:145–148.
bes2-9638227	Eickwort, G.C., Eickwort, J.M., Gordon J. & Eickwort, M. A. Solitary behavior in a high-altitude population of the social sweat bee *Halictus rubicundus* (Hymenoptera: Halictidae). Behav Ecol Sociobiol (1996) 38:227–233.
bes2-9639293	Roces, F. & Hölldobler, B. Use of stridulation in foraging leaf-cutting ants: mechanical support during cutting or short-range recruitment signal? Behav Ecol Sociobiol (1996) 39:293–299.
bes2-9639381	Keasar, T., Shmida, A. & Motro, U. Innate movement rules in foraging bees: flight distances are affected by recent rewards and are correlated with choice of flower type. Behav Ecol Sociobiol (1996) 39:381–388.
bes2-9741151	Trumbo, S.T., Huang, Zhi-Yong & Robinson, G.E. Division of labor between undertaker specialists and other middle-aged workers in honey bee colonies. Behav Ecol Sociobiol (1997) 41:151-163.
bes2-9842009	van Baaren, J. & Boivin, G. Learning affects host discrimination behavior in a parasitoid wasp. Behav Ecol Sociobiol (1998) 42:9-16.
bes2-9842239	Heinze, J., Hölldobler, B. & Yamauchi, K. Male competition in Cardiocondyla ants. Behav Ecol Sociobiol (1998) 42:239-246.
bes2-9843067	Beye, M., Neumann, P., Chapuisat, M., Pamilo, P. & Moritz, R.F.A. Nestmate recognition and the genetic relatedness of nests in the ant *Formica pratensis*. Behav Ecol Sociobiol (1998) 43:67-72.
bes2-9844193	Pankiw, T., Page Jr, R.E. & Fondrk, M.K. Brood pheromone stimulates pollen foraging in honey bees (*Apis mellifera*). Behav Ecol Sociobiol (1998) 44:193-198
bes2-9945047	Agrawal, A.A. & Dubin-Thaler, B.J. Induced responses to herbivory in the Neotropical ant-plant association between *Azteca* ants and *Cecropia* trees: response of ants to potential inducing cues. Behav Ecol Sociobiol (1999) 45:47-54.
bes2-9945177	Bee, M.A., Perrill, S.A. & Owen, P.C. Size assessment in simulated territorial encounters between male green frogs (*Rana clamitans*). Behav Ecol Sociobiol (1999) 45:177-184.
bes2-9946171	Fewell, J.H. & Bertram, S.M. Division of labor in a dynamic environment: response by honeybees (*Apis mellifera*) to graded changes in colony pollen stores. Behav Ecol Sociobiol (1999) 46:171-179.

[+] available at http://link.springer.de/link/service/journals/00265/bibs/

Appendix III (continued)

Table A3-1c. List of Articles[+] in Sub-corpus oec1

oec1-97109265	Traugott, M.S. & Stamp, N.E. Effects of chlorogenic acid- and tomatine-fed caterpillars on performance of an insect predator. Oecologia (1997) 109:265–272.
oec1-97110143	Hubbard, J. A. & McPherson, G.R. Acorn selection by Mexican jays: a test of a tri-trophic symbiotic relationship hypothesis. Oecologia (1997) 110:143-146.
oec1-97111209	Wehling, W.F. & Thompson, J.N. Evolutionary conservatism of oviposition preference in a widespread polyphagous insect herbivore, *Papilio zelicaon*. Oecologia (1997) 111:209-215.
oec1-97111570	Grostal, P. & Walter, D.E. Kleptoparasites or commensals? Effects of *Argyrodes antipodianus* (Araneae: Theridiidae) on *Nephila plumipes*(Araneae: Tetragnathidae). Oecologia (1997) 111:570-574.
oec1-97112566	Hawkins, B.A. & Marino, P.C. The colonization of native phytophagous insects in North America by exotic parasitoids. Oecologia (1997) 112:566-571.
oec1-97112572	Yeargan, K.V. & Quate, L.W. Adult male bolas spiders retain juvenile hunting tactics. Oecologia (1997) 112:572-576.
oec1-98114343	Hoffmann, J.H. & Moran, V.C. The population dynamics of an introduced tree, *Sesbania punicea*, in South Africa, in response to long-term damage caused by different combinations of three species of biological control agents. Oecologia (1998) 114:343-348.
oec1-98115154	Bauer, G. Structure and function of a non-interactive, reactive insect-plant system. Oecologia (1998) 115:154-160.
oec1-98115184	Morse, D.H. The effect of wounds on desiccation of prey: implications for a predator with extra-oral digestion. Oecologia (1998) 115:184-187.
oec1-98115427	Wright, M.G. & Samways, M.J. Insect species richness tracking plant species richness in a diverse flora: gall-insects in the Cape Floristic Region, South Africa. Oecologia (1998) 115:427-433.
oec1-98115434	Stiefel, V.L. & Margolies, D.C. Is host plant choice by a clytrine leaf beetle mediated through interactions with the ant *Crematogaster lineolata*? Oecologia (1998) 115:434-438.
oec1-99120252	Morse, D.H. Choice of hunting site as a consequence of experience in late-instar crab spiders. Oecologia (1999) 120:252-257.
oec1-99120268	Karban, R. & Kittelson, P.M. Effects of genetic structure of *Lupinus arboreus* and previous herbivory on *Platyprepia virginalis* caterpillars. Oecologia (1999) 120:268-273.
oec1-99120274	Hopkins, R.J. & Ekbom, B. The pollen beetle, *Meligethes aeneus*, changes egg production rate to match host quality. Oecologia (1999) 120:274-278.
oec1-99120304	Burger, J.C., Patten, M.A., Rotenberry, J.T. & Redak, R.A. Foraging ecology of the California gnatcatcher deduced from fecal samples. Oecologia (1999) 120:304-310.

[+] available at http://link.springer.de/link/service/journals/00442/bibs/

Appendix III (continued)

Table A3-1d. List of Articles[+] in Sub-corpus oec2

oec2-97109313	Halaj, J., Ross, D.W. & Moldenke, A.R. Negative effects of ant foraging on spiders in Douglas-fir canopies. Oecologia (1997) 109:313–322.
oec2-97109454	Eubanks, M.D., Nesci, K.A., Petersen, M.K., Liu, Z. & Sanchez, H.B. The exploitation of an ant-defended host plant by a shelter-building herbivore. Oecologia (1997) 109:454-460.
oec2-97110360	Fritz, R.S., McDonough, S.E. & Rhoads, A.G. Effects of plant hybridization on herbivore-parasitoid interactions. Oecologia (1997) 110:360-367.
oec2-97110539	Stange, G. Effects of changes in atmospheric carbon dioxide on the location of hosts by the moth, *Cactoblastis cactorum*. Oecologia (1997) 110:539-545.
oec2-97112081	Keese, M.C. Does escape to enemy-free space explain host specialization in two closely related leaf-feeding beetles (Coleoptera: Chrysomelidae)? Oecologia (1997) 112:81-86.
oec2-97112209	Gaume, L., McKey, D. & Anstett, M.-C. Benefits conferred by "timid" ants: active anti-herbivore protection of the rainforest tree *Leonardoxa* africanaby the minute ant *Petalomyrmex phylax*. Oecologia (1997) 112:209-216.
oec2-97117258	Krupa, J.J. & Sih, A. Fishing spiders, green sunfish, and a stream-dwelling water strider: male-female conflict and prey responses to single versus multiple predator environments. Oecologia (1998) 117:258-265.
oec2-98113391	Frenzel, M. & Brandl, R. Diversity and composition of phytophagous insect guilds on Brassicaceae. Oecologia (1998) 113:391-399.
oec2-98114382	Desouhant, E., Debouzie, D. & Menu, F. Oviposition pattern of phytophagous insects: on the importance of host population heterogeneity. Oecologia (1998) 114:382-388.
oec2-98117133	Erelli, M.C., Ayres, M.P. & Eaton, G.K. Altitudinal patterns in host suitability for forest insects. Oecologia (1998) 117:133-142.
oec2-98117420	Orr, M.R. & Seike, S.H. Parasitoids deter foraging by Argentine ants (*Linepithema humile*) in their native habitat in Brazil. Oecologia (1998) 117:420-425.
oec2-99118166	Müller, C. & Hilker, M. Unexpected reactions of a generalist predator towards defensive devices of cassidine larvae (Coleoptera, Chrysomelidae). Oecologia (1999) 118:166-172.
oec2-99118381	Russell, R.W. Precipitation scrubbing of aerial plankton: inferences from bird behavior. Oecologia (1999) 118:381-387.
oec2-99119191	Toft, S.& Wise, D.H. Growth, development, and survival of a generalist predator fed single- and mixed-species diets of different quality. Oecologia (1999) 119:191-197.
oec2-99120437	Sword, G.A. & Dopman, E.B. Developmental specialization and geographic structure of host plant use in a polyphagous grasshopper, *Schistocerca emarginata* (= *lineata*) (Orthoptera: Acrididae). Oecologia (1999) 120:437-445.

[+] available at http://link.springer.de/link/service/journals/00442/bibs/)

Appendix IV

Table A4-1. Distribution of % Sentences Selected for Abstracting per Document

Section	Percentage of sentences selected per section by document [+]								
	0	1-10	11-20	21-30	31-40	41-50	51-60	61-70	71-75
Introduction	1	2	8	14	8	19[*]	3	1	1
Method	18[**]	17	10	8	2	2	-	-	-
Results	3	7	13	20	8	4	2	-	-
Discussion	1	3	9	19	11	9	4	-	1
Total	23	29	40	61	29	34	9	1	2

[+] no. of sentences selected per document: average = 17; lowest = 8, highest = 26; eight documents selected 21 sentences as the basis for abstracting.

[*] For 19 documents, i.e. 1/3 of corpus, 41-50% of selected sentences came from Introduction section.

[**] For 18 documents, i.e. about 1/3 of corpus, no sentence were selected from the Method section for abstracting.

Table A4-2. Distribution of Sentences: (a) in Corpus, and (b) Selected for Abstracting

Sentences	Introduction	Method	Results	Discussion	Total
Full text	1316 (16.58%)	2524 (31.80%)	1647 (20.75%)	2451 (30.88%)	7938 (100.01%)
Selected	330 (33.88%)	106 (10.88%)	234 (24.02%)	304 (31.21%)	974 (99.99%)
Reduction Factor (RF)	4:1	24:1	7:1	8:1	8:1
Abstract					534

RF = no. of sentences in full text / no. selected sentences

Table A4-3. Percentage of Sentences with Five Words or Less

Sub-corpus	Number of sentences			%
	With 4 wds	With 5 wds	In article	
bes1	1	4	1903	0.26
bes2	2	4	1937	0.31
oec1	9	6	1675	0.90
oec2	2	4	2423	0.25
Total	14	18	7938	0.40

Appendix IV (continued)

Table A4-4. Sentences Cued[+] and Not Cued Selected by Author for Abstracting

sub-corpora	Number of sentence				Total
	selected by author		not selected by author		
	with cues	without cues	with cues	without cues	
bes1	60	161	264	1418	1903
bes2	58	167	258	1454	1937
oec1	48	200	239	1188	1675
oec2	76	204	363	1780	2423
Total	242 (3 %)	732 (9 %)	1124 (14 %)	5840 (74 %)	7938

[+] With any of the following lexemes: CONCLUDE, INDICATE, RESULT, SHOW, STUDY, SUGGEST.

Table A4-5. Initial/Final Position Sentences Cued and Not Cued Selected by for Abstracting

	Section				Total I + R + D (excluding M)
	I	M	R	D	
Init/Fin + Cued + Selected	47	2	20	49	116
Init/Fin + Cued	130	91	106	235	471
Init/Fin + Selected	161	52	109	136	406
Init/Fin	516	918	686	928	2130

Init/Fin = Initial or Final sentence; Selected = sentence selected for abstracting by an author;
Cued = sentence cued with any of the following lexemes: CONCLUDE, INDICATE, RESULT, SHOW, STUDY, SUGGEST;

In our study, the trait in question was not an artificial addition (e.g., colored leg bands), and manipulation of asymmetry in the trait was (in part) within the range of natural variation. [D-3-4; bes1-9945087]

Our selection study showed that there was significant negative selection on male balloon volume and a trend towards positive selection on male body size. [D-2-2; bes1-9945161]

The results of this study indicate that per capita *T. coweni* foundress survival and net biomass production decline with an increase in mean number of foundresses per gall. [D-11-2; bes1-9843095]

To conclude, large males moved more, suggesting that size may have a role in sexual selection in this species, but experiments are needed to determine the relative importance of mass-independent drumming rate and mass-related mobility. [D-5-1; bes1-9946123]

This study indicates the nature of the cost-benefit ratio for an insect predator. [D-7-1; oec1-99120252]

In this study we have shown that individual *M. aeneus* match oviposition rate to the changes in available oviposition resource. [D-6-1; oec1-99120274]

Fig. A4-1. Cued^{+} Sentences Not Selected for Abstracting
($^{+}$ The cues included any of the following lexemes:
CONCLUDE, INDICATE, RESULT, SHOW, STUDY and SUGGEST)

Appendix V

Table A5-1. Type II Substitutions Involving Major Parts of Speech

Verb	
(A5-1)	Johnson *et al.* proposed a tri-trophic symbiotic relationship [I-4-4]
→	it has been suggested that a tri-trophic relationship [A-1-4; oec1-97110143]
	Syn(propose) = suggest (WN)
(A5-2)	we examine differentiation [I-3-1]
→	We analyzed geographic differentiation [A-1-1; oec1-97111209]
	Syn(examine) = analyze (WN)
Adjective	
(A5-3)	*A. antipodianus* cause reduced weight gain … [I-4-4]
→	…decreased weight gain may have resulted from … by *A. antipodianus*. [A-1-4; oec1-97111570]
	Syn(reduced) → decreased (WN)
(A5-4)	were a small component of the diet of California gnatcatchers [D-2-6]
→	were only minor components of the gnatcatcher diet [A-1-7; oec1-99120304]
	Syn(small) = minor (WN)
(A5-5)	initiated foraging at younger ages [D-3-4]
→	initiated foraging at earlier ages [A-1-7; bes2-9741151]
	Syn(young) = early (WN)
Adverb	
(A5-6)	one of the most broadly distributed butterflies [M-7-3]
→	one of the most widely distributed … butterflies [A-1-1; oec1-97111209]
	Syn(broadly) = widely (WN)
(A5-7)	the high degrees of resource utilization usually observed. [D-8-1]
→	the frequently observed high degree of the resource utilization by the insect. [A-1-8; oec1-98115154]
	Syn(usual) = frequent (¬WN), but **Syn**(usual) = common (WN); **Syn**(common) = frequent (WN)
(A5-8)	An early experiment indicated that tussock caterpillars forced to feed on branches that … grew less rapidly … [I-5-3]
→	Previous experiments … suggested that feeding caused by either of these two folivores could reduce … [A-1-2; oec1-99120268]
	Syn(early) = previous (WN); **Syn**(indicate) = suggest (WN)

Appendix V (continued)

Table A5-2a. Unlikely Type II Domain Substitutes: Special Lexical Resource

$\mathbf{A_0}$(leaf)	= foliar
Hypernym(activity)	= behavior
Hypernym(parasitism)	= activity
Hypernym(caterpillar)	= folivore
Hypernym(congener)	= species
Hypernym(control)	= experiment
Hypernym(foreleg)	= leg
Hypernym(psychodid)	= fly
Hypernym(treatment)	= experiment
Juven(weevil)	= larva
Magn(discriminate)	= actively
Magn(feed)	= intensively, actively
Mult(honeybee)	= colony
Sing(nest)	= ant
Syn(endemic)	= restricted [to]
Syn(epigeic$^+$)	= non-web-building
Syn(radiation)	= diversification
Syn(vector)	= agent
$\mathbf{S_{res}}$(autotomize)	= loss

$^+$ Epigean = living near or on the ground surface, applied specifically to insects.

Appendix VI

Table A6-1. Examples of Illocution Markers Deleted

(A6-1) ~~Our results positively demonstrate that~~ hexane-extractable compounds associated with brood stimulate pollen foraging. [D-5-1]

→ $\phi_{illocution\_marker}$ Hexane extracts of larvae containing brood pheromone stimulated pollen foraging. [A-1-4; bes2-9844193]

(A6-2) ~~These results indicate that~~ males modify their vocal behavior in different ways or to different degrees depending on the frequency of an opponent's call. [D-2-7]

→ $\phi_{illocution\_marker}$ the frequency of an opponent's calls elicits a differential modification of calling behavior, … [A-1-7; bes2-9945177]

(A6-3) ~~The most important finding of the present study was that~~ male drumming rate was positively related to male survival in both field and laboratory conditions, suggesting that male drumming activity may be a reliable indicator of male viability. [D-1-1]

→ $\phi_{illocution\_marker}$ Males drumming at the highest rate survived better than males drumming at a lower rate in both laboratory and field conditions. [A-1-5; bes1-9946123]

(A6-4) ~~As demonstrated in previous studies,~~ when faced with predation risk in the open water where fish occur, male water striders shifted to predator-free microhabitats … [D-2-3]

→ $\phi_{illocution\_marker}$ In the presence of both predators, male water strider behavior (microhabitat use and activity) … was … [A-1-11; oec2-97117258]

(A6-5) ~~Thus~~, ~~our objective was~~ to assess foraging ecology… through identification of prey … [I-1-4]

→ $\phi_{connective}$ $\phi_{illocution\_marker}$ We identified arthropod fragments … to gain insight into its foraging ecology … [A-1-3; oec1-99120304]

Table A6-2: Occurrence of modifier *significant* deleted, retained or inserted during abstracting.

Sub-corpus	Deleted	Retained	Inserted
bes1	4	13	11
bes2	-	12	5
oec1	6	16	-
oec2	11	17	4
Total	21	58	20

Appendix VII

(a) Mary sold tomatoes on Monday.
Mary purchased cars on Tuesday.
...
→ Mary and John each did business all week.

[from Dalianis (1999:386)]

(b) He heard an explosion.
He phoned the police.
→ He heard an explosion and (he) phoned the police.
or → When he heard an explosion, he phoned the police.

[from Leech & Svartvik (1975:288)]

Fig. A7-1. Examples of aggregation from: (a) text generation, (b) grammar books

Appendix VIII

Document VIII-1

Full text (oec2-97117258)	Abstract
T1@: Many studies have experimentally examined$_z$ how the risk of predation affects$_x$ prey habitat use, activity, foraging behavior$_y$ ~~and group dynamics~~. [I-1-1]	A1: Many studies have experimentally addressed$_z$ the effects of$_x$ a particular predator species on prey behavior$_y$. [A-1-1] (I-1-1) → (A-1-1)
T2a#: ~~Moreover,~~ most prey species face multiple predators. [I-1-6] T2b#: An adaptive response by a prey to one predator might increase its exposure to another predator, and the risk to prey$_x$ can vary as these predators interact. [I-1-7]	A2a: In nature, however, prey frequently face multiple species of predators A2b: that often vary in their predatory mode and in their level of predation risk$_x$. [A-1-2] (I-1-6) + (I-1-7) → (A-1-2)
T3@: ~~Although many studies have looked at prey escape and avoidance behaviors,~~ only a few$_x$ experimental studies have addressed$_y$ the effects of predation risk on mating dynamics. [I-2-1] T3@: ~~To our knowledge,~~ no$_x$ experimental study has considered$_y$ how prey mating dynamics are affected by multiple predators. [I-2-3]	A3: Relatively few$_x$ studies have considered$_y$ prey responses under these complex conditions. [A-1-3] (I-2-3) → (A-1-3) (I-2-1) = (I-2-3)

= ft-sentence that is a partial match for ab-sentence; @ = ft-sentence that is a full match for ab-sentence;

Document VIII-1 (continued)

Full text (oec2-97117258)	Abstract
T4a#: ~~Recent work has shown that~~ the mating system of the stream-dwelling water strider, *Aquarius remigis*$_x$, can shift adaptively with changing social and ecological conditions. [I-3-1]	A4a: In Kentucky, the stream-dwelling water strider (*Aquarius remigis*)$_x$ coexists with many potentially dangerous predators$_y$,
T4b#: In particular, the microhabitat distribution, general activity, mating activity, and patterns of non-random mating of *A. remigis*$_x$ are all strongly influenced by the presence of the predatory green sunfish (*Lepomis cyanellus*). [I-3-2]	A4b: two of which are the green sunfish (*Lepomis cyanellus*) and the fishing spider (*Dolomedes vittatus*). [A-1-]
T4b: Water striders., however, face other potential predators$_y$ besides green sunfish. [I-4-1]	(I-3-1) + (I-3-2) + (I-4-1) + (I-4-2) + (I-4-5) → (A-1-4)
T4b#: These include backswimmers (Notonectidae), fishing spiders *Dolomedes triton*), green frogs *Rana clamitans*), and brown trout *Salmo trutta*)$_y$. [I-4-2]	
T4a: In the south-eastern United States, all four of these potential predators$_y$ can coexist with *A. remigis*$_x$. [I-4-3]	
T4b#: Here we compared: (1) the effects of green sunfish and fishing spiders *Dolomedes vittatus*) on male and female *A. remigis*$_x$ microhabitat distribution, general activity, aggressive behavior, mating activity, mating frequency, mating duration, and mortality; and (2) the effect that sunfish and spiders together have on these same behavioral variables. [I-4-5]	
T4: In Kentucky, *A. remigis*$_x$ begins breeding in February or March, depending on weather conditions, and continues until late May or early June. [M-1-3]	
T5: In small Kentucky streams, green sunfish are one of the most potentially dangerous fish predators. [M-3-2]	A5: Green sunfish occupy stream pools and attack water striders from below. [A-1-5]
T5@: Sunfish attack water striders from below in deeper water, ~~while fishing spiders perch vertically on rocks and overhanging vegetation along the shore where they may catch and lift water striders off the water's surface.~~ [M-5-2]	(M-5-2) → (A-1-5)

= ft-sentence that is a partial match for ab-sentence; @ = ft-sentence that is a full match for ab-sentence;

Document VIII-1 (continued)

Full text (oec2-97117258)	Abstract
T6@: Sunfish attack water striders$_y$ from below in deeper water, while fishing spiders perch vertically on rocks and overhanging vegetation along the shore where they may catch and lift water striders$_y$ off the water's surface$_x$. [M-5-2] !! good example of one ft sentence going to two ab-sentence. !! (M-5-2) → (A-1-5) + (A-1-6)	A6: In contrast, fishing spiders hunt along stream shorelines where they perch on overhanging vegetation or rocks and attack water striders$_y$ near shore$_x$. [A-1-6] (M-5-2) → (A-1-6)
T7@: ~~Here~~ we compared: ~~(1)~~ the effects of$_z$ green sunfish and fishing spiders *Dolomedes vittatus*)$_y$ on male and female *A. remigis*$_x$ microhabitat distribution, ~~general activity, aggressive behavior, mating activity, mating frequency, mating duration, and mortality; and (2) the effect that sunfish and spiders together have on these same behavioral variables.~~ [I-4-5]	A6: We compared how *A. remigis* individuals$_x$ respond to$_z$ these two very different predators$_y$ in pools with one or both predators. [A-1-7] (I-4-5) → (A-1-7)
T8: Within each pool, water striders had three sources of potential refuge$_z$: they could: ~~(1)~~ climb onto styrofoam blocks; ~~(2)~~ climb up the walls of the tank; or (3) sit$_y$ just out of the water on the edge of the downstream riffle. [M-8-2] T8#: The presence of fish caused$_x$ male water striders to increase their use of three types of refuge$_z$ (riffles; climbing out of the water; and staying$_y$ on the water but near the edges of pools), and to decrease activity. [R-3-3] !! staying$_y$ → sitting$_y$; T8#: Associated with these shifts in microhabitat$_z$ use and activity, fish presence caused$_x$ a reduction in the number of aggressive males on the water. [R-3-4] T8: For activity, fewer males were active and aggressive in fish pools and fish+spider pools than in spider pools or predator-free pools. [R-4-4]	A8: The presence of sunfish in pools had strong effects on$_x$ male water strider behavior, including increased use of three types of refuge$_z$ from sunfish (riffles, climbing out of the water, sitting$_y$ on the water but at the edges of pools), decreased activity A8: and a decreased number of aggressive males on the water. [A-1-8] (R-3-3) and (R-3-4) → (A-1-8)

= ft-sentence that is a partial match for ab-sentence; @ = ft-sentence that is a full match for ab-sentence;
!! = notes/comments;

Document VIII-1 (continued)

Full text (oec2-97117258)	Abstract
T9@: ~~In contrast,~~ spiders significantly influenced only one aspect of male activity or microhabitat use$_x$; male water striders avoided spiders by shifting away from the edges of pools ~~(where spiders were most dangerous)~~. [R-3-5]	A9: Spiders also influenced water strider behavior$_x$; male water striders avoided spiders by shifting away from the edges of pools. [A-1-9] (R-3-5) → (A-1-9)
T10b#: ~~Specifically,~~ more$_z$ males were in riffles and along the edges, in fish pools$_x$ than in spider pools$_y$, but fish pools$_x$ did not differ significantly from pools with fish+spiders. [R-4-3] T10a,b#: Comparing fish pools$_x$ and spider pools$_y$ in this study, male water striders exhibited different$_z$ anti-predator responses that appear adaptively associated with the specific predator species. [D-2-2]	A10a: Comparisons of the effects of the two predator species showed that in general, A10b: anti-predator responses by male water striders were stronger$_z$ in pools with fish alone$_x$ than in those with spiders alone$_y$. [A-1-10] (R-4-3) + (D-2-2) → (A-1-10)
T11b#: ~~Specifically,~~ more males were in riffles and along the edges, in fish pools than in spider pools, but fish pools did not differ significantly$_x$ from pools with fish+spiders$_y$. [R-4-3] T11a#: ~~As demonstrated in previous studies,~~ when faced with predation risk in the open water where fish occur, male water striders shifted to predator-free microhabitats (riffles, out of water, edges of pools) and reduced activity that should reduce conspicuousness to predators. [D-2-3]	A11a: In the presence of both predators$_y$, male water strider behavior (microhabitat use and activity) A11b: was generally similar$_x$ to behavior in the presence of fish alone. [A-1-11] (D-2-3) + (R-4-3) → (A-1-11)

= ft-sentence that is a partial match for ab-sentence; @ = ft-sentence that is a full match for ab-sentence;

Document VIII-1 (continued)

Full text (oec2-97117258)	Abstract
T12b#: In contrast, female water strider behavior was quite different from male behavior, and females showed relatively little$_y$ response$_x$ to predators$_z$. [R-5-1] T12a#: Fish had no significant$_y$ impact$_x$ on the behavior of female water striders, and spiders caused only a decrease in the proportion of females out of the water. [R-5-3]	A12a: In contrast, female water striders showed no significant$_y$ response$_x$ to the presence of sunfish, A12b: and little$_y$ response to the presence of spiders$_z$. [A-1-12] (R-5-1) + (R-5-3) → (A-1-12)
T13#: Females tended to be relatively inactive$_z$, spending most of their time$_y$ out of water or in riffles. [D-3-2] T13#: For example, taking refuge in riffles reduced female exposure to both males and predators, while inactivity helps females avoid drawing the attention of both males$_y$ and predators. [D-3-5]	A13: This lack of response$_z$ could be because females spent much of their time$_y$ in refuges even in the absence of predators (apparently hiding from harassment by males$_y$). [A-1-13] (D-3-2) + (D-3-5) → (A-1-13)
T14: In the presence of this predator, adult *A. remigis* reduce their general activity and shift their microhabitat use away from the center of pools. [M-3-3] T14: Green sunfish also cause a reduction in water strider mating activity. [M-3-4] T14@: A two-way ANOVA indicated that both fish and spiders caused decreases in water strider mating activity. [R-6-1]	A14: Both spiders and fish caused decreases in water strider mating activity. [A-1-14] (R-6-1) → (A-1-14) !! Results reported in Method.
T15@: The presence of fish caused decreases$_x$ in both mating frequency$_x$ and mating duration, while spiders caused a significant reduction in mating duration, but not mating frequency. [R-6-3] !! sentence (R-6-3) is split into two sentences (A-1-15) and (A-1-16);	A15: The presence of fish reduced$_x$ both the number of matings per pool (mating frequency)$_x$, and mean mating durations. [A-1-15] (R-6-2) → (A-1-15)

= ft-sentence that is a partial match for ab-sentence; @ = ft-sentence that is a full match for ab-sentence;
!! = notes/comments;

Document VIII-1 (continued)

Full text (oec2-97117258)	Abstract
T16@: ~~The presence of fish caused decreases in both mating frequency and mating duration, while~~ spiders caused a significant reduction in mating duration, but not mating frequency. [R-6-3]	A16: Spiders induced a decrease in mean mating duration, but not in mating frequency. [A-1-16] (R-6-3) → (A-1-16)
T17#: ~~Tukey's test showed that~~ the largest reductions in mating activity occurred in the presence of both predators. [R-6-4]	A17: The largest reductions in mating activity occurred in pools with both predators present. [A-1-17] (R-6-4) → (A-1-17)
T18#: Thus the two predators together should cause greater mortality than would be expected based on a simple summing of the isolated effects of the two. [D-4-5] T18#: With either sunfish or spiders present, water striders spent about 30% of their time in tandem. [D-5-10]	A18: Pools with either spiders or fish alone suffered 15-20% water strider mortality during our experiment (versus no mortality in predator-free pools). [A-1-18] (D-4-5) + (D-5-10) → (A-1-17) !! 15-20% : not in full text;
T19@: In the presence of both predators, water striders have no microhabitat refuge; avoidance of one predator increases exposure to the other. [D-4-4]	A19: Extant theory suggests that when prey face conflicting microhabitat responses to two predators (as in this study), the predators should have facilitative effects on predation rates (i.e., prey that avoid one predator are often killed by the other and vice versa. [A-1-19] (D-4-4) → (A-1-19)

= ft-sentence that is a partial match for ab-sentence; @ = ft-sentence that is a full match for ab-sentence;
!! = notes/comments;

Document VIII-1 (continued)

Full text (oec2-97117258)	Abstract
T20: Mortality rates in the two-predator pools, were, ~~if anything,~~ lower (though not statistically significantly lower) than expected, based on the multiplicative risk model. [D-4-8]	A20: Mortality rates in pools with both predators present, however, were not significantly different from that predicted, by a null model of multiple predator effects. [A-1-20] (D-4-8) → (A-1-20)
T21: What might explain this lack of risk enhancement? [D-5-1-8) T21#: ~~Sih et al. suggested that~~ risk enhancement in the presence of multiple predators can be prevented if prey also use other compensatory defenses that have generalized effects (i.e., that reduce predation ~~risk from~~ both predators). [D-5-2] T21#: ~~As expected,~~ water striders exhibited a drastic reduction in mating activity when exposed to multiple predators. [D-5-9] T21#: This, along with reductions in general activity, probably explained the lack of risk enhancement in this multiple predator system. [D-5-12]	A21: The lack of predator facilitation can be explained by the compensatory reductions in water strider activity and mating activity in the presence of both predators. [A-1-21] (D-5-2) + (D-5-9) and (D-5-12) → (A-1-21)

= ft-sentence that is a partial match for ab-sentence; @ = ft-sentence that is a full match for ab-sentence;

Document VIII-2

Full text (oec2-98117420)	Abstract
T1#: The Argentine ant, *Linepithema humile* (formerly *Iridomyrmex humilis*), is one of the most widespread and destructive invasive ants in the world. [I-1-3] T1#: Its invasions threaten endemic arthropods in Hawaii and eliminate native ants in California$_x$, Australia, and South Africa. [I-1-4]	A1: The Argentine ant, *Linepithema humile*, has invaded sites across Africa, Australia, Europe, and North America$_x$. [A-1-1] (I-1-3) + (I-1-4) → (A-1-1) !!Europe not in full text;
T2#: Its invasions threaten endemic arthropods in Hawaii and eliminate native ants in California, Australia, and South Africa. [I-1-4] T2#: Argentine ants also tend homopterans and augment their destructiveness in agriculture. [I-1-6] !! domain knowledge used here;	A2: In its introduced ranges it eliminates native ants and tends agricultural pests. [A-1-2] (I-1-4) + (I-1-6) → (A-1-2)
T3: If Argentine$_x$ ants are ever to be controlled using natural enemies, the ecology of the ant in South America$_x$ requires further study. [I-3-5]	A3: Few studies have examined the ecology of Argentine$_x$ ants in their native habitat. [A-1-3] (I-3-5) → (A-1-3)
T4#: Here we examined whether Argentine ant foraging in Brazil is suppressed by the presence of *Pseudacteon* parasitoids. [I-4-1] !! personification: we examined → this study examined	A4: This study examined the effects of parasitoid flies, genus *Pseudacteon*, on the foraging behavior of Argentine ants in part of their native range in southern Brazil. [A-1-4] (I-4-1) → (A-1-4)

= ft-sentence that is a partial match for ab-sentence; @ = ft-sentence that is a full match for ab-sentence; !! = notes/comments;

Document VIII-2 (continued)

Full text (oec2-98117420)	Abstract
T5#: *Pseudacteon* parasitoids were active during daylight hours at temperatures above 18°C. [R-4-1] !! active ⇒ attack; domain knowledge required; T5#: ~~During our observations,~~ none of the many other ant species interacting with Argentine ants was attacked by *P. pusillum.* [D-3-5]	A5: *Pseudacteon* parasitoids commonly attacked Argentine ants, but not other ant species, in daylight at temperatures above 18°C. [A-1-5] (R-4-1) + (D-3-5) → (A-1-5)
T6: ~~Overall,~~ parasitoids appeared to be more important than temperature in inhibiting the exploitation of food resources by Argentine ants in Brazil. [I-4-4] T6: The arrival of parasitoids at a foraging trail caused Argentine ants to abandon recruitment and, in most instances, abandon the resource. [R-6-1] T6#: By retreating underground, Argentine ants avoided high rates of parasitoid attack. [R-6-3] T6#: At all four sites, Argentine ants ~~completely~~ abandoned resources in the presence of parasitoids. [D-1-2]	A6: Argentine ants abandoned food resources and returned underground in the presence of parasitoids. [A-1-6] (R-6-3) + (D-1-2) → (A-1-6)
T7@: By retreating underground, Argentine ants avoided high rates of parasitoid attack. [R-6-3] !! Avoid high → diminish; T7@: Their tendency to return underground appears to allow them to avoid high rates of parasitism. [D-1-3]	A6: Parasitoid attack rates diminished as Argentine ants retreated underground. [A-1-7] (R-6-3) → (A-1-7) (R-6-3) = (D-1-3)

= ft-sentence that is a partial match for ab-sentence; @ = ft-sentence that is a full match for ab-sentence;
!! = notes/comments;

Document VIII-2 (continued)

Full text (oec2-98117420)	Abstract
T8: We predicted that if parasitoids suppress foraging, Argentine ant abundance at food resources would be lowest when and where parasitoids were active$_w$. [M-3-2] T8#: Considering first instances without parasitoids$_x$, there was some tendency for ants to be more abundant at night or early morning than during the day$_y$, although these differences were not significant. [R-1-2] T8: The one exception to this pattern was CJ in October, when parasitoids were scarce$_x$ and Argentine ants were abundant at baits. [R-1-4] !! Reasoning for (R-1-2) and (R-1-4); T8#: Considering the abundance of Argentine ants at food resources in the absence of *Pseudacteon* parasitoids$_x$, and their scarcity in the presence of parasitoids, it is likely that parasitoids influence exploitative competition among Argentine ants and other ants for food. [D-3-7] !! absence ⇒ inactivity; world knowledge required;	A8: Where parasitoids were present$_w$, Argentine ants were abundant at food resources only during times of day$_y$ when parasitoids were inactive$_x$. [A-1-8] (R-1-2) + (D-3-7) → (A-1-8)
T9#: Considering first instances without parasitoids, there was some tendency for ants to be more abundant at night or early morning than during the day, although these differences were not significant. [R-1-2] T9#: The one exception to this pattern was CJ in October, when parasitoids were scarce and Argentine ants were abundant at baits. [R-1-4] T9: Another potential application based on the results presented here might be in Agriculture, where Argentine ants protecting homoptera conceivably could be inhibited during the daytime by the introduction of *Pseudacteon* parasitoids. [D-3-10]	A9: Where parasitoids were absent, Argentine ants were abundant at food resources throughout the day. [A-1-9] (R-1-2) + (R-1-4) → (A-1-9) difficult

= ft-sentence that is a partial match for ab-sentence; @ = ft-sentence that is a full match for ab-sentence;

Document VIII-2 (continued)

Full text (oec2-98117420)	Abstract
T10#: Overall, parasitoids appeared to be more important than temperature in inhibiting the exploitation of food resources by Argentine ants in Brazil. [I-4-4] !! exploitation of food resources → foraging; !! !! Results reported in Introduction; T10#: Overall, there was no indication that Argentine ants could not tolerate foraging at all but the very highest temperature reported in Fig.1, although high temperatures clearly had a suppressive effect on foraging. [R-3-3]	A10: Overall, the presence of parasitoids explained observed variation in Argentine ant foraging far better than temperature, although temperature had some effect. [A-1-10] (I-4-4) + (R-3-3) → (A-1-10)
T11@: Considering the abundance of Argentine ants at food resources in the absence of *Pseudacteon* parasitoids, and their scarcity in the presence of parasitoids, it is likely that parasitoids influence exploitative competition among Argentine ants and other ants for food. [D-3-7] T11: Another potential application based on the results presented here might be in Agriculture, where Argentine ants protecting homoptera conceivably could be inhibited during the daytime by the introduction of *Pseudacteon* parasitoids. [D-3-10]	A11: The results suggest that *Pseudacteon* parasitoids inhibit the ability of Argentine ants to gather food resources in their native habitat in Brazil. [A-1-11] (D-3-7) → (A-1-11)

= ft-sentence that is a partial match for ab-sentence; @ = ft-sentence that is a full match for ab-sentence; !! = notes/comments;

Document VIII-3

Full text (oec1-97111209)	Abstract
T1#: In the anise swallowtail butterfly, *Papilio zelicaon* Lucas, however, although some populations differ in their use of hosts, individuals within and among most populations differ little in how they rank potential host plants for oviposition preference. [I-1-7] T1#: The geographic structure of oviposition preference in *P. zelicaon* is of particular interest, because it is one of the most broadly distributed butterflies in western North America. [I-2-2] T1#: *P. zelicaon* is also one of the most polyphagous butterflies, feeding on at least 65 species in 29 genera of Umbelliferae and four species in three genera of Rutaceae. [I-2-5] T1#: Here we examine differentiation in oviposition preference among 13 populations of *P. zelicaon* in the Pacific Northwest of North America. [I-3-1] !! (I-1-7): difficult to read "differ ... differ"; to determine subject	A1: We analyzed geographic differentiation in oviposition preference in the anise swallowtail butterfly, *Papilio zelicaon* Lucas, which is one of the most widely distributed and polyphagous butterflies in western North America. [A-1-1] (I-2-2) + (I-3-1) + (I-1-7) + (I-2-5) → (A-1-1) !! no "geographic differentiation" in full text;
T2#: Here we examine differentiation in oviposition preference among 13 populations of *P. zelicaon* in the Pacific Northwest of North America. [I-3-1] T2#: Our purpose is to test the hypothesis resulting from an earlier study that oviposition preference in *P. zelicaon* is highly conserved over broad geographic areas, despite major differences in local availability of host plant species. [I-3-4] T2#: Overall, 1200 km separate the two most distant populations in this study. [D-1-6] !! much repetition;	A2: Among 13 populations that span 1200 km of the range of *P. zelicaon* in the Pacific Northwest of North America, the overall oviposition preference hierarchy has not diverged significantly, even though these populations differ in the plant species they use in the field. [A-1-2] (I-3-1) + (I-3-4) + (D-1-6) → (A-1-2)

= ft-sentence that is a partial match for ab-sentence; @ = ft-sentence that is a full match for ab-sentence; !! = notes/comments;

Document VIII-3 (continued)

Full text (oec1-97111209)	Abstract
T3@: The oviposition preference hierarchy exhibited among populations has not undergone major reorganization, even when these populations differ locally in the hosts they have available. [D-1-7]	A3: The results indicate that differences in host availability and use have not favored major reorganizations in the preference hierarchy of ovipositing females. [A-1-3] (D-1-7) → (A-1-3)
T4#: The oviposition preference hierarchy exhibited among populations has not undergone major reorganization, even when these populations differ locally in the hosts they have available. [D-1-7]	A4: Instead, this butterfly has a conserved, preference hierarchy that varies within a narrow range among populations. [A-1-4] (D-1-7) → (A-1-4)
T5#: The 13 populations did not differ in oviposition preference for these four plant species. [R-1-4] T5@: The oviposition preference hierarchy exhibited among populations has not undergone major reorganization, even when these populations differ locally in the hosts they have available. [D-1-7] T5#: The comparison among these four species, however, suggests that relative ranking does not vary much, at least among some common hosts of this butterfly species. [D-2-5]	A5: All populations ranked the four test plant species in the same overall relative order, even though these populations differ in the plant species they use in the field. [A-1-5] (R-1-4) + (D-2-5) + (D-1-7) → (A-1-5)

= ft-sentence that is a partial match for ab-sentence; @ = ft-sentence that is a full match for ab-sentence; !! = notes/comments;

Document VIII-4

Full text (oec1-98115184)	Abstract
T1a#: Desiccation is a particularly important concern for predators that inject their prey with proteolytic enzymes, with resulting breakdown of the prey's body tissues outside the predator's digestive system (extra-oral digestion). [I-1-5] T1b#: A major source of difficulty for predators that employ extra-oral digestion is that the prey's increasing viscosity decreases the predator's ability to extract this resource from the carcasses. [D-1-1]	A1a: Predators that inject prey with proteolytic enzymes, thereby breaking down their tissues for subsequent ingestion, A1b: run the risk that desiccation will hinder eventual retrieval of resources from these prey. [A-1-1] (I-1-5) + (D-1-1) → (A-1-1)
T2@: ~~For those that retain their prey intact,~~ wounds made in capturing the prey ~~and subsequently feeding on it~~ could exacerbate the rate of desiccation, ~~especially since they often change their feeding sites on prey after the initial kill~~. [I-2-3]	A2: Wounds made in capture might exacerbate this problem. [A-1-2] (I-2-3) → (A-1-2)
T3a#: For this analysis I used small hover flies *Toxomerus marginatus* (Syrphidae) killed by crab spiders *Misumena vatia* (Thomisidae). [I-3-1] T3a#: Spider-killed and chill-killed flies also lost mass in the sun at a similar rate, which was highly significantly more rapid than the rate of those placed in the shade, approaching dry mass at 3 h. [R-1-3] !! 'M' 'fly' = {chill-killed, intact fly, unwounded fly}; T3b#: Both groups lost about 1/12 of their wet body mass over 3 h, the amount of time that spiders of this size usually feed on *Toxomerus* and about 1/6 of their wet mass over 7 h. [R-1-2] T3c#: Thus, individuals in sun and shade lost mass at strikingly different rates, but wounds did not affect this rate of loss. [R-1-4]	A3a: However, desiccation rates of small syrphid flies *Toxomerus marginatus* (Diptera: Syrphidae) killed by juvenile crab spiders *Misumena vatia* (Araneae: Thomisidae) and intact dead syrphid flies did not differ A3b: over the normal period of feeding, A3c: though desiccation rates in shade and sun differed several-fold. [A-1-3] (I-3-1) + (R-1-2) + (R-1-3) + (R-1-4) → (A-1-3)

= ft-sentence that is a partial match for ab-sentence; @ = ft-sentence that is a full match for ab-sentence; !! = notes/comments;

Document VIII-4 (continued)

Full text (oec1-98115184)	Abstract
T4a#: The mass of the spider inflicting the wound did not affect the loss in mass of the fly in either sun or shade. [R-2-2] T4b#: Location of the carcass (shade or sun) greatly affected rate of body loss, but individuals killed by spiders did not differ in rate of loss from those killed by chilling in either shade or sun, strongly suggesting that the wounds produced by the spiders did not themselves represent a significant source of evaporative loss. [D-3-3]	A4a: Neither the size of the spider (and presumably the size of the wounds it inflicted) A4b: nor the location of the wounds on the flies' bodies affected desiccation rates. [A-1-4] (R-2-2) + (D-3-3) → (A-1-4)
T5#: Predators that attack extremely large prey, relative to their own body size, present an extreme in processing costs. [I-1-2] T5#: However, the size of the spider (and accompanying chelicerae) generating the wound did not have a significant effect, also suggesting that the wounds were insignificant factors in the loss of body mass. [D-3-6]	A5: Thus, this tactic of prey handling does not exact an added processing cost on *Misumena*. [A-1-5] (I-1-2) + (D-3-6) → (A-1-5)

= ft-sentence that is a partial match for ab-sentence; @ = ft-sentence that is a full match for ab-sentence;

Glossary

ADJUNCT(-IVAL) A term used in grammatical theory to refer to an optional or secondary element in a construction. An adjunct may be removed without the structural identity to the rest of the construction being affected. The term may be given a highly restricted sense, as when it is used in Quirk Grammar to refer to a subclass of adverbials. (Crystal, 1997)

ADVERBIAL Any phrase in a sentence which is functionally similar to an adverb in that it modifies the action in respect of time, manner, place or circumstance. (Trask, 1997)

APPOSITION A traditional term retained in some models of grammatical description for a sequence of units which are constituents at the same grammatical level, and which have an identity or similarity of reference. In *John Smith, the butcher, came in*, for example, there are two noun phrases; they have identity of reference, and they have the same syntactic function (as indicated by the omissibility of either, without affecting the sentence's acceptability. (Crystal, 1997)

CATENATIVE A term used in some grammatical descriptions of the verb phrase to refer to a lexical verb which governs the non-finite form of another lexical verb, as in one possible analysis of *she likes to go, she wants to see, she hates waiting*, etc. In generative grammar, such constructions are known as control and raising constructions. (Crystal, 1997)

CO-HYPONYM See hyponym. (Crystal, 1997)

COMPLEMENTIZER A word which introduces a complement clause, such as that or whether. (Trask, 1997)

COMPOUND WORD A word made up of two or more other words. This exhibits a kind of covert syntax based mainly on prepositional phrases: the compound *teapot* can be paraphrased only as 'a pot for tea', not 'a pot of tea'. Innumerable semantic relationship of this kind occur among compounds, some easy to interpret in isolation, others dependent on context. *London goods*, for example, may be 'goods in London', 'goods for London', 'goods from London'. Paraphrasing is not, however, always straightforward, even when the context is clear. What paraphrase is best for *steamboat*: 'a boat that uses steam', 'a boat using steam', ... Precise paraphrase is impossible, but imprecise paraphrases still work adequately, because the relation between *steam* and *boat* is clear enough. (McArthur, 1992)

COMPLEMENT A term used in the analysis of grammatical function to refer to a major constituent of sentence or clause structure, traditionally associated with 'completing' the action specified by the verb. In generative grammar, a complement is a sister constituent of a zero-level category. Categories other than the verb are also sometimes said to take complements, e.g. *a student of physics*. (Crystal, 1997)

COMPLEX PREPOSITION Preposition consisting of more than one word. (Leech & Svartvik, 1977)

CONNECTIVE A term used in the grammatical classification of words to characterize words or morphemes whose function is primarily to link linguistic units at any level. Conjunctions are the most obvious types (e.g. *and, or, while, because*), but several types of adverb can be seen as connective ('conjuncts' such as *therefore, however, nevertheless*), as can some verbs (the copulas *be, seem*, etc). (Crystal, 1997)

CONTENT WORDS or CONTENTIVES Words which have stateable lexical meaning. Alternative terms include LEXICAL and FULL WORDS. (Crystal, 1997)

CONVERSE A term often used in semantics to refer to a sense relation between lexical items. Converse terms display a type of oppositeness of meaning, illustrated by such pairs as bug/sell, parent/child, employer/employee, and above/below. *Buy* is the converse of *sell* and vice versa. In such a relationship found especially in the definition of reciprocal social roles, spatial relationship, and so on, there is an interdependence of meaning, such that one member of he pair presupposes the other member. In this respect, 'converseness' contrast with complementarity,

where there is no such symmetry of dependence, and with the technical sense of antonymy, where there is a gradation between the opposite. (Crystal, 1997)

COORDINATE Coordinate terms are words that have the same hypernym. (WordNet 1.6)

DELETION A basic operation within the framework of transformational grammar, which eliminates a constituent of an input phrase-marker. (Crystal, 1997)

ELLIPSIS The omission of one or more elements that can be recovered, understood, from the linguistic or situational context. (Huddleston, 1988)

FUNCTION The relationship between a linguistic form and other parts of the linguistic pattern or system in which it is used. In grammar, e.g. the noun phrase can 'function' in clause structure as subject, object, complement, etc., these roles being defined distributionally. (Crystal, 1997)

GENERIC A lexical stem or proposition which refers to a class of entities, e.g. *the bat is an interesting creature, bats are horrid, ...* (Crystal, 1997)

HEDGE An application in pragmatics and discourse analysis of a general sense of the word ('to be non-committal or evasive') to a range of items which express a notion of imprecision or qualification. Examples include *sort of, more or less, I mean, approximately, roughly*. Hedges may also be used in combination: *something of the order of 10 per cent, more or less*. (Crystal, 1997)

HOLONYM The name of the whole of which the meronym names a part. Y is a holonym of X if X is a part of Y. (WordNet 1.6)

HYPERNYM See *hyponym*.

HYPERONYM See *hypernym*.

HYPONYM A term used in semantics as part of the study of the sense relations which relate lexical items. 'Hyponymy' is the relationship which obtains between specific and general lexical items, such that the former is 'included' in the latter (i.e. 'is a hyponym of' the latter). For example, a *cat* is a hyponym of *animal*, *flute* of *instrument*. In each case, there is a superordinate term (sometimes called a HYPERNYM or HYPERONYM), with respect to which the subordinate term can be defined, as in the usual practice in dictionary definitions ('a cat is a type of animal ...'). The set of terms which are hyponyms of the same superordinate term are CO-HYPONYMS. (Crystal, 1997)

INSERTION A basic syntactic operation within the framework of transformational grammar which introduces a new structural element into a string; ... 'lexical insertion' which inserts lexical items at particular places in grammatical structure. (Crystal, 1997)

INVERSION A term used in grammatical analysis to refer to the process or result of syntactic change in which a specific sequence of constituent is seen as the reverse of another. (Crystal, 1997)

LEXEME A term used by some linguists to refer to the minimal *distinctive unit* in the semantic system of a language. The lexeme is thus postulated as the abstract unit underlying such sets of grammatical variants as *walk, walks, walking, walked*, or *big, bigger, biggest*. ... (Crystal, 1997)

MERONYM(Y) A term used in semantics as part of the study of the sense relations which relate lexical items. 'Meronymy' is the relationship which obtains between 'parts' and 'wholes', such as *wheel* and *car*, or *leg* and *knee*. 'X is a part of Y' contrasts especially with 'X is a kind of Y' relationship (hyponymy). (Crystal, 1997)

METADISCOURSE "Writing about writing, whatever does not refer to the subject matter being addressed" (Williams, 1981). An author usually writes at two levels. At one level, propositional content on a subject is supplied, and at another, metadiscourse which does not contribute to propositional content, but helps a reader "organize, classify, interpret, evaluate, and react to [the propositional] material", is added. (Vande Kopple, 1985)

METONYMY A figure of speech. The name of an attribute of an entity is used in place of the entity itself: the bottle (for the drinking of alcohol) or the violins (in The second violins are playing well). (Crystal, 1997)

MODAL A term used in grammatical and semantic analysis to refer to contrasts in mood signalled by the verb and associated categories. In English, modal contrasts are primarily expressed by a subclass of auxiliary verbs, e.g. may, will, can. Modal verbs share a set of morphological and syntactic properties which distinguish them from the other auxiliaries, e.g. no -s, -ing, or -en forms. (Crystal, 1997)

MODIFICATION A term for the dependence of one grammatical unit on another, the less dependent unit being delimited or made more specific by the more dependent unit: the adjective *good* modifying the noun *weather* in the phrase *good weather*, the noun *diamond* modifying the noun *mines* in *diamond mines*, the adverb *strikingly* modifying the adjective *handsome* in *strikingly handsome*. A distinction is made between PRE-MODIFICATION (modifying by preceding) and POST-MODIFICATION (modifying by following). Clauses may also be modifiers in phrases, usually post-modifiers of nouns, such as the relative clause in 'the bag *that you are carrying*'. The dependence of a subordinate clause on its superordinate clause is generally not described in terms of modification: the subordinate clause in 'I know *that you are there*' is not said to be a modifier. Some grammarians, however, use the term *sentence modifier* for adverbials (including adverbial clauses) that express a comment on the sentence or clause: *fortunately* in '*Fortunately*, no one was hurt'; *in all probability* in '*In all probability*, it is closed by now'; the *since*-clause in '*Since you're here*, you may as well make yourself useful.' Although the distinction is obvious between such examples and clear instances of adverbials functioning as modifiers of verbs (such as 'The band is playing *too loudly*', there is no agreement on how to draw the line between sentence modifiers and verb modifiers or on how many relational categories to establish for adverbials. (McArthur, 1992)

NOMINALIZATION Nominalization refers to the process of forming a noun from some other word class (e.g. red+ness) or (in classical transformational grammar especially) the derivation of a noun phrase from an underlying clause (e.g. *Her answering of the letter ...* from *She answered the letter*). (Crystal, 1997)

PARADIGMATIC A basic term for the set of substitutional relationships a linguistic unit has with other units in a specific context. (Crystal, 1997)

PARAPHRASE A term used in linguistics for the result or process of producing alternative versions of a sentence or text without changing the meaning. One sentence may have several paraphrases, e.g. *The dog is eating a bone, A bone is being eaten by the dog, It's the dog who is eating a bone*, and so on. (Crystal, 1997)

PERSONIFICATION Reference to something general or abstract as if it were an individual: e.g. love is personified in *love conquers all.* (Matthews, 1997)

PHRASAL VERB Verbs may form combinations with adverbial particles which, in their form and behaviour ar like prepositional adverbs., for example *He's applied for a new job. Her parents strongly objected to her travelling alone*. The noun phrase following the proposition is termed the prepositional object. (Leech & Svartvik, 1975)

PREPOSITIONAL VERB A verb may also form a combination with a preposition., for example *He's applied for a new job. Her parents strongly objected to her travelling alone*. The noun phrase following the proposition is termed the prepositional object. (Leech & Svartvik, 1975)

REDUCTION Term usually refers to a clause (a reduced clause) which lacks one or more of the elements required to enable it to be used as a full, independent construction, e.g. *to see the book.* Such clauses may be referred to as 'abbreviated', elliptical or contracted, but different approaches often introduce distinctions between these terms. Other units are sometimes referred to as 'reduced', such as phrases (phone's ringing) [instead of telephone] and words (e.g. it's him) [is → 's]. (Crystal, 1997)

SENSE A meaning of a word in WordNet. Each sense of a word is in a different synset. (WordNet 1.6)

SUBLANGUAGE Term coined by Harris (1968) to describe a subset of sentences in a language which can be generated from a special set of grammatical rules, some of which belong to the grammar of the language, others of which are unique to the sublanguage itself. Thus, in the sublanguage of an aviation hydraulics maintenance manual *the*-deletion is required: *Depressuric ϕ hydraulic system.* Sublanguage are also characterize by constraints on collocations. For example, in the sublanguage of stock market reports, intransitive verbs of motion (e.g. *plunge, drop*) are combined only with certain nouns and certain adverbs, while this same combinations are not found in the standard language: *Mines plunge sharply. The gold index dropped sharply.* (Bussmann, 1996)

SUBSTITUTION A term used in linguistics to refer to the process or result of replacing one item by another at a particular place in a structure. In grammar, the structural context within which this replacement occurs is known as a substitution frame, e.g. The __ is angry, and the set of items which can be used paradigmatically at its given place is know as substitution class. A word which refers back to a previously occurring element of structure may be called a substitute word. (Crystal, 1997)

SUPERORDINATE See HYPERNYM.

SUPPLETION (SUPPLETIVE) A term used in morphology to refer to cases where it is not possible to show a relationship between morphemes through a general rule, because the forms involved have different roots. A suppletive is the grammar's use of an unrelated form (i.e. with a different root) to complete a paradigm as in the present-pat relation of *go~went*, or in the comparative form *better* in relation to *good.* (Crystal, 1997)

SYNSET A synonym set; a set of words that are interchangeable in some context. (WordNet 1.6)

SYNTAGMATIC A fundamental term to refer to the sequential character of speech, a string of constituents (often) in linear order. (Crystal, 1997)

SYNTACTIC FUNCTION The grammatical role of units within the construction immediately containing them. (Huddleston, 1988)

TROPONYM A verb expressing a specific manner elaboration of another verb. X is a troponym of Y if to X is to Y in some manner. (WordNet 1.6)