Dear Author pdf/**A/2**

Thank you for publishing with Taylor & Francis. A PDF proof of your article is attached for **urgent proofreading** along with an 'Agreement for the Transfer of Copyright'. Please read the following instructions carefully.

RETURNING YOUR PROOFS/CORRECTIONS
(1) Please print out the article, check it carefully, and attend to any queries listed on the proof query sheet at the end of the article. Amendments should be marked clearly on to the hard copy in ink (**do not mark the PDF file**).

(2) Please send corrections back by fax or urgent post to the address below within **72 hours** of receipt. When faxing, marks must be clear and legible in all instances and not too close to the margin in case they are omitted during transmission. Alternatively, you may send your corrections by email (quoting Journal Title, page and line number of each correction), to the email address proof-queries@tandf.co.uk, ensuring all corrections are clear and concise. (Please keep a copy of all the corrections as a backup). Please find attached a copy of some commonly used proofreading marks.

AGREEMENT FOR THE TRANSFER OF COPYRIGHT
Please print out, complete and sign the attached 'Agreement for the Transfer of Copyright' form. This should be returned to the address below. (If you have already signed and returned a copyright form by prior arrangement, then please disregard this request. However, you should be aware that non-receipt of a completed form could possibly delay publication of your article).

Address for return of the proofs/signed copyright form:
'Journal Title (**not** Article Title)'
Floor 2, Journals Production
Taylor & Francis Ltd
4 Park Square
Milton Park, Abingdon
Oxfordshire OX14 4RN, UK

Tel: +44 (0)1235 828600
Fax: +44 (0)1235 829000 or +44(0)1235 829009
Email: proof-queries@tandf.co.uk

E-PRINTS (Electronic Offprints) / OFFPRINTS
An E-print (PDF file, from which you may print or distribute no more than 50 copies to friends or colleagues) may be sent to you, via email, up to two weeks prior to publication, and a copy of the issue in which the article appears will be sent by standard mail. Alternatively, you may receive 50 printed offprints (if you wish to take this option, you should advise us by email). If you wish to order additional offprints or copies of the journal issue, then please contact the Production Editor named on the 'cc' list on this email.

Many thanks in advance for your co-operation. If you have any queries, please contact us at: proof-queries@tandf.co.uk

http://www.tandf.co.uk
http://www.taylorandfrancis.com

If you have problems accessing the PDF file, please access the Adobe website and download the free Acrobat Reader from: Http://www.adobe.com/products/acrobat/readstep.html

Taylor & Francis
Taylor & Francis Group plc

Production Editorial Department, Taylor & Francis Ltd.
4 Park Square, Milton Park, Abingdon OX14 4RN
Telephone: +44 (0) 1235 828600; Facsimile: +44 (0) 1235 829000

PROOF CORRECTION MARKS

Instruction		Mark in text	Mark in margin
Leave unchanged		~~text~~	(stet)
Extraneous marks or damaged letter		encircle	×
Delete		/ or /—/	ꝗ
Insert or replace text		λ	/text to be added
Add (λ) or substitute (/):	full stop	λ /	⊙
	decimal point	λ /	⊙
	comma	λ /	,/
	semi-colon	λ /	;/
	colon	λ /	(:)
	apostrophe or quotation mark	λ /	ʼ ʼ
	superscript	λ /	^{2}
	subscript	λ /	$_{2}$
	hyphen	λ /	/-/
	short or long rule	λ /	/en/ or /em/
	oblique	λ /	(/)
Wrong typeface or size		encircle	wf
Change to:	roman (upright)	encircle	(rom)
	italic	underline	(ital)
	capital letters	underline three times	(caps)
	small capitals	underline twice	(s.c)
	bold type	wavy underline	(bold)
	lower case letters	encircle	(l.c.)
	Greek letters	encircle	(gk) adding Greek letter
Delete and close up		co~~p~~py	ꝗ
Reduce space		in / copy	less #
Close up space		in co⁀py	⁀
Insert space		inλcopy	#
Make space in line equal		⊥	eq #
Insert space between lines		)————(	
Reduce space between lines		(————)	
New paragraph		[	(para)
Run on, no new paragraph		⊂	(run on)
Transpose letters or words		⊔⊓	⊔⊓
Transpose lines		═══	
Take character to next line		⊏	(take over)
Take character to previous line		⊐	(take back)
Raise text on page		⌐——¬	(raise)
Lower text on page		⌙——⌟	(lower)
Check vertical alignment		‖	‖
Check horizontal alignment		═	═

Informatics-Assisted Statistical Analysis of Vocabularies

Tünde-Molnár Lengyel, *Hungary*

Abstract

As one of the most important concerns in the training of library and informatics-experts, information provision requires a thorough knowledge of professional texts used in different fields. This is, however, due to the sheer number of publications appearing in various periodicals, which is a difficult goal to accomplish when making a thorough study, if the information conveyed there is virtually impossible to find. Consequently, the importance of research abstracts has significantly increased as they allow the reader or the librarian to study the most relevant sections of given articles or essays in a fraction of the time otherwise required in reading the whole work. Moreover, with the help of this method, overlaps or redundancies are eliminated. Furthermore, the increasing number of publications poses a difficult challenge for the documentation preparation experts as well.

L'analyse statistique aidée par informatique des vocabulaires

Un des principaux problèmes dans la formation des bibliothécaires et des experts informaticiens collectant des informations exige une connaissance approfondie des textes professionnels utilisées dans divers domaines. Ceci est dû au nombre considérable de publications qui proviennent de multiples périodiques, ce qu'est un but difficile à atteindre qu'on fait une étude approfondie, si l'information fournie est virtuellement impossible à trouver. En conséquences l'importance des abstracts de recherches à augmenté de facon significative car ils permettent au lecteur ou un bibliothécaire d'étudier les sections les plus importantes des articles ou essais donnés dans une fraction du temps qui autrement serait recu par lire l'ensemble du document. De plus, avec l'aide de cette méthode, les recherches et les interférences sont éliminés. Par ailleurs, le nombre croissant de publications crée un challenge difficile aussi bien pour les experts préparant la documentation.

EDV-unterstützte statistische Analyse von Vokabularen

Als eines der wichtigsten Themen in der Ausbildung von Bibliotheks- und Informationsexperten erfordert die Bereitstellung von Informationen eine gründliche Kenntnis von in unterschiedlichen Bereichen verwendeten professionellen Texten. Das Erstellen einer gründlichen Studie ist jedoch allein schon wegen der reinen Anzahl der in verschiedenen Zeitschriften erscheinenden Veröffentlichungen ein schwieriges Unterfangen, weil die dort verfügbare Information praktisch nicht überblickt werden kann. Folglich ist die Bedeutung kurzer Zusammenfassungen (Abstracts) erheblich gestiegen, da sie dem Leser oder dem Bibliothekar erlauben, die relevantesten Abschnitte eines Artikels oder einer Abhandlung in einem Bruchteil der sonst für das Lesen der ganzen Arbeit erforderlichen Zeit zu studieren. Außerdem werden mit Hilfe dieser Methode Überlappungen oder Redundanzen vermieden. Ohnehin stellt die wachsende Anzahl von Veröffentlichungen eine schwierige Herausforderung für die Dokumentationsexperten dar.

In the following section, I would like to examine the possibilities and limitations of the automation of research abstract preparation, in addition to surveying the procedures and technological implementations facilitating the compilation of research abstracts in the Hungarian language.

Being well informed – both in one's professional and everyday life – is one of modern life's most important values, and libraries play a crucial role in the realization of this goal. While librarians performing an information provision function are expected to be well-versed in the texts of different disciplines, the rapid development of each field makes it difficult to keep one's knowledge current and up to date. No matter which discipline one focuses on, a thorough examination of the relevant research results published in periodicals is virtually impossible.

Education Media International
ISSN 0952-3987 print/ISSN 1469-5790 online © 2003 International Council for Educational Media
http://www.tandf.co.uk/journals
DOI: 10.1080/0952398032000105085

The requirements of one's profession or the expectations of society also lead to an increase in the number of articles since the evaluation and potential advancement of a college instructor or academic researcher in any discipline are based upon the number of articles published and lectures presented at scholarly conferences. Subsequently, concurrently with the growing number of publications, the article's novelty content declines and becomes redundant along with overlaps also increasing. Therefore, the role of research abstracts becomes increasingly significant since they help the reader to save a tenth of the time in obtaining relevant, or (theoretically) resources; being the most important information compared to a full reading of a given text. Furthermore, the compilation of research abstracts also helps in the elimination of redundancies.

While, in my opinion, the significance and importance of research abstracts are beyond any dispute, no one can be expected to read through all the periodicals containing research abstracts and each respective excerpt. However, research abstracts allow one to gain more information relevant to his or her field; another important benefit is that the reading of these abstracts helps one to decide whether the full reading of a given article is necessary.

So far, we have examined the importance of research abstracts from a user's point of view and, even after a brief glance, we can safely conclude that the overview of the latter poses a significant challenge in most disciplines. When we look at the other side of the coin, the compilation of research abstracts, we have to contend with difficulties there too. Ever proliferating publications make the work of abstract and documentation compilers increasingly arduous because traditional methods cannot yield a comprehensive or approximate reproduction of the given materials in the respective fields. A greater reliance on the computer in the work of documentation appears to be the only solution.More and more procedures should be elaborated to facilitate the presentation of the most important elements of articles, or possibly books.

Below, we will survey the possibilities and limitations of automated abstract compilation, in addition to examining the methods and technological implementations relevant to the computer-assisted abstract preparation process.

A potential crisis in the documentation process

Attempts to find a solution to the aforementioned problems have not been made in the past few years, as the term 'documentation preparation crisis' had already appeared in Hungary in 1963. Sándor Szalay's book, describing the contemporary state of mechanized abstract preparation, also listed some interesting statistical data proving that the proliferation of documents started in the 1800s: 'In 1750 – 12, in 1800 more than 90, in 1850 more than 900, in 1900 approximately 9000, and in 1950 more than 80,000 scholarly publications (periodicals and other cyclical materials) were produced world wide in the field of natural science' (Szalay, 1963, p. 5). Furthermore, according to contemporary or recent data, the number of periodicals published globally in 2000 was around 160,000, offering additional proof that publication figures have increased exponentially thus the term is pertinent in 1963 and has gained an even higher relevance by now!

The abstract preparation process

The term 'abstract' functions as an umbrella concept for a procedure aimed at the reproduction and re-presentation of the content of a given document. Justified by its optimal suitability for computer-assisted processing, computer-assisted abstract preparation procedure will be examined in greater detail.

An abstract is a short-form reproduction of any information communication, either written or verbal. Abstract preparation takes place when the essence of a statement or communication is reproduced in the author's own words, thi is known as homotopic communication. Indicative communication (one referring to a given subject) or the enumeration of the components of the communication or statement is another form of abstract preparation. The preparation of indicative communication can be automated as there is no need for an expert to interpret the spoken or read material and highlight its most important parts in a form intelligible for everyone; one 'merely' needs to enumerate the most important elements. (These two types often appear in a combined form.)

The above classification follows a content-based approach concerning the given material. In treating the concept of abstract from the point of view of linguistics or logic, the methods described above are called summarizing (summa) or selective (excerptum) procedures. In the case of the summarizing abstracts, the preparer reproduces sections or parts that he or she considers important with his own words, and with selective abstracts the description of the unchanged form of the textual components or units constitute the abstract. (Szalai, 1963, p. 9)

Methods

Several statistical methods are used exclusively for specialised analysis, and there are those that are indispensable for the compilation of abstracts. Frequency analysis belongs to the latter category. The first step for the automation of abstract preparation is ascertaining how many times certain words (treated as separate units) appear in the text. Then the given set of data is organized according to frequency, and thus the statistical verbal rendering of the text is obtained.

Frequency analysis

Zipf discovered a certain regularity in the distribution of the words and structures of the text. He examined Joyce's *Ulysses* and, having arranged the words of the novel according to occurrence, he asserted 'that the product of the cumulative occurrence figures and the inclusive frequency values is constant.' (Horváth and Papp, 1999, p.107).

In order to perform a frequency analysis in the text, the roots of the words of the text (type) along with the different forms of occurrence (sign) have to be identified and the occurrence will be ranked according to frequency. Whereas the identification of roots is a painstakingly long effort calling for the inclusion of the computer in the work process, in the case of Hungarian texts this is the most difficult task. One possible solution could be computer-assisted linguistics, which started in Hungary in 1960 with the introduction of mechanized translations. This period was primarily characterized by the elaboration of the foundations of the mechanized translation algorithm from Russian to Hungarian language. The second era (1967–71) is defined by the work of documentary linguistics experts elaborating a syntactic analysis method of their own. The third lexicologic phase (1972–78) responding to the needs of literary critics or philologists included the development of software for language teaching and the compilation of quantitative-analysis based frequency dictionaries focusing on colloquial and literary Hungarian. However, these research results were so closely associated with certain scholars that the disbanding of the Documentation Group in 1972 in Budapest all but eliminated linguistics research efforts. With 1979, the fourth stage began with experiencing an attempt made to make up for the loss of research findings in the 1970s. Because of a dynamic development of language processing systems throughout Europe, Hungarian researchers elaborate the MI language culminating in the arrival of a Hungarian morphological analysis application method.

The appearance of personal software in the 1990s led to rapid developments. The elaboration of a spelling and grammar check system taking into consideration of the characteristics of the Hungarian language was a significant achievement of this period. In this system, the composition of words – that is, the connection between the root and the suffix – was described by an algorithm. Morphologic – the firm responsible for its development – became one of the leading companies in the computerized linguistics field after Microsoft purchased its programme. The newer versions of this software examine the context and can eliminate the irrelevant interpretations (Prószéki, 1989, pp. 489–492). Today, more Hungarian institutions achieve world-wide reputation due to their computerized linguistics efforts. For example, the ILP (Inductive Logic Programming) developed in the Artificial Intelligence Research Laboratory of the Hungarian Academy of Science and of Szeged University played a pioneering role in the introduction of experimental linguistic applications.

While the above-mentioned results help in the identification of roots of words in the Hungarian language, they have not yet been applied in the field of library science. Following the identification of roots, the counting of the words contained by a text is necessary for the frequency analyses, a task that can be carried out via simple programming commands. In order to perform a frequency analysis or prepare an abstract or excerpt, the significant expressions contained by the given text have to be identified.

Following Zipf's laws, significant expressions constitute the given domain of the frequency list which is dependent upon the respective discipline. However, it is true, in all disciplines, that these expressions do not constitute the beginning or the end of the given list. The list of significant words is obtained when the Gauss curve – defined according to the empirical method characteristic of the given discipline – is projected on to the frequency distribution function (Horváth and Papp, 1999, p. 56). As far as Hungarian texts are concerned, few disciplines devised a frequency dictionary facilitating the construction of the Gauss curve. Presently, the Hungarian Academy of Science is involved in the compilation of word frequency dictionaries.

According to Luhn's notion, elaborated in 1951, the multiple occurrences of certain doublets or triple word constructs can be helpful in computer-assisted identification of terms carrying relevant information. Having omitted trivial expressions, the weighting of adjacent words and triple word constructs help to identify relevant sections of the text. Consequently, non-trivial expressions containing two or more units get a higher weighting value than their counterparts appearing only once in the text. Having established the weighting process, one has to define which units he or she wants to retrieve as relevant location, either in the form of a sentence or a paragraph. Subsequently, the automation process starts when a numerical value is assigned to the chosen unit, based upon the weighting process and the sentences and paragraphs reflecting the highest numerical value are retrieved as a result.

Difficulties encountered during the mechanical abstract preparation process

- Since trivial expressions separate doublets or triple word constructs carrying relevant information, the identification of the latter can be problematic. However, with the help of special software designed for this purpose, this problem can be solved in cases where expressions separated by a few words in the text.
- While the identification of the subject of the given sentence is done via a reference to person(s) or event(s) listed in the previous sentences, this process can be made easier by textual analysis.
- To achieve more precise results, the statistical analysis should be expanded along with the recognition of the first occurrence of significant words and the assignment of weighting of commensurate value.
- The expansion of the focus of inquiry from the words to the position of a sentence within the paragraph is likely to provide more accurate results since authors tend to introduce their subject at the beginning, and summarize their findings in the closing sentence of the given paragraph. Since these sentences often contain relevant information, they appear to deserve a higher weighting value and the paragraphs could be analysed from the same vantage point as well.
- A writer's attempt to enliven the text with diverse expressions denoting the same concept or person significantly decreases the efficiency of this process.

Summary

In conclusion, I would point out that this procedure can only be used in case of discursive texts during which the author concentrates on one topic and uses objective statements with a consistent vocabulary, form, and structure instead of writing in a literary language. Generally, we can conclude that the Luhn method is more effective in case of scientific statements, publications and reports, than it would be in case of a sophisticated literary work. Whereas automated abstract preparation systems have not yet been implemented in Hungary, the demand for such apparatus appears to increase.

References

Antal, L (1975) *The Foundations of Content Analysis*, Tömegkommunikációs Kutatóközpont, Budapest.

Holsti, OR (1969) *Content Analysis for the Social Sciences and humanities*, Addison Wesley, Reading, MA.

Horváth, T and Papp, I (1999) *Librarians' Reference Book 1*, Osiris compendia, Budapest.

Horváth, T and Papp, I (2001) *Librarians' Reference Book 2*, Osiris compendia, Budapest.

Krippendorff, K (1980) *Content Analysis: An Introduction to Its Methodology,* Sage Publications, Beverly Hills, CA.

Pietil, V (1973) *Sisallön Erittely*, Gaudeamus, Helsinki.

Prószéky, G (1989) *Computer-Assisted Linguistics*, Számítástechnika-alkalmazási Vállalat, Budapest

Spiegel, MR (1988) *Theory and Problems of Statistics*, McGraw-Hill, Berkshire.

Stone, PJ, Dunphy, DC, Smith, MS and Ogilvie, MG (1966) *The General Inquirer, A Computer-Approach to Content Analysis in the Behavioral Sciences*, MIT Press, Cambridge, MA.

Biographical note

Lengyelné Molnár Tünde, is an assistant professor at Eszterházy Károly College, Hungary. Presently she is enrolled at the Library informatics PhD programme at Eötvös Loránd University. The title of her doctoral research and thesis is 'Vocabulary statistical analysis and text condensation methods relevant to the automated abstract preparation process'. She has been actively involved in the research efforts of the Multimedia Research Laboratory of Eszterházy Károly College where, in addition to fulfilling her instructional tasks, she prepares multimedia presentations on her own.

Address for correspondence

Lengyelné Molnár Tünde, Eszterházy Károly College, Department of Informatics, 3300 Eger Eszterházy tér 1, Hungary; e-mail: mtunde@ektf.hu

Annotations from REMI100163.pdf

Page 2

Annotation 1; Label: Author Query; Date: 25/6/2003 10:45:33
AQ 1 – Page 212
Szalay (or Szalai?) (1963) not in refs. Please provide details or delete citation

Page 4

Annotation 1; Label: Author Query; Date: 25/6/2003 10:46:31
AQ 2 – Page 214
The following references are not cited in main text. Please cite or delete from refs:
Antal (1975)
Holsti (1969)
Horváth, T and Papp, I (2001)
Krippendorff, K (1980)
Pietil, V (1973)
Spiegel, MR (1988)
Stone et al (1966)

AN AGREEMENT FOR THE TRANSFER OF COPYRIGHT

IN RELATION TO THE CONTRIBUTION OF YOUR ARTICLE ('THE ARTICLE') ENTITLED:

…………………………………………………………………………………………………….

BY:……………………………………………………………………………………………………

WHICH WILL BE PUBLISHED IN R-EMI

…………………………………………………………………………………………………….

In order to ensure both the widest dissemination and protection of material published in our Journal, we ask authors to assign the rights of copyright in the articles they contribute. This enables International Council for Educational Media ('us' or 'we') to ensure protection against infringement. In consideration of the publication of your Article, you agree to the following:

1. You assign to us with full title guarantee all rights of copyright and related rights in your Article. So that there is no doubt, this assignment includes the assignment of the right to publish the Article in all forms, including electronic and digital forms, for the full legal term of the copyright and any extension or renewals. Electronic form shall include, but not be limited to, microfiche, CD-ROM and in a form accessible via on-line electronic networks. You shall retain the right to use the substance of the above work in future works, including lectures, press releases and reviews, provided that you acknowledge its prior publication in the journal.

2. Our publisher, Taylor & Francis Ltd, shall prepare and publish your Article in the Journal. We reserve the right to make such editorial changes as may be necessary to make the Article suitable for publication, or as we reasonably consider necessary to avoid infringing third party rights or law; and we reserve the right not to proceed with publication for whatever reason.

3. You hereby assert your moral rights to be identified as the author of the Article according to the UK Copyright Designs & Patents Act 1988.

4. You warrant that you have at your expense secured the necessary written permission from the appropriate copyright owner or authorities for the reproduction in the Article and the Journal of any text, illustration, or other material. You warrant that, apart from any such third party copyright material included in the Article, the Article is your original work, and does not infringe the intellectual property rights of any other person or entity and cannot be construed as plagiarising any other published work. You further warrant that the Article has not been previously assigned or licensed by you to any third party and you will undertake that it will not be published elsewhere without our written consent.

5. In addition you warrant that the Article contains no statement that is abusive, defamatory, libelous, obscene, fraudulent, nor in any way infringes the rights of others, nor is in any other way unlawful or in violation of applicable laws.

6. You warrant that wherever possible and appropriate, any patient, client or participant in any research or clinical experiment or study who is mentioned in the Article has given consent to the inclusion of material pertaining to themselves, and that they acknowledge that they cannot be identified via the Article and that you will not identify them in any way.

7. You warrant that you shall include in the text of the Article appropriate warnings concerning any particular hazards that may be involved in carrying out experiments or procedures described in the Article or involved in instructions, materials, or formulae in the Article, and shall mention explicitly relevant safety precautions, and give, if an accepted code of practice is relevant, a reference to the relevant standard or code.

8. You shall keep us and our affiliates indemnified in full against all loss, damages, injury, costs and expenses (including legal and other professional fees and expenses) awarded against or incurred or paid by us as a result of your breach of the warranties given in this agreement.

9. You undertake that you will include in the text of the Article an appropriate statement should you have a financial interest or benefit arising from the direct applications of your research.

10. If the Article was prepared jointly with other authors, you warrant that you have been authorised by all co-authors to sign this Agreement on their behalf, and to agree on their behalf the order of names in the publication of the Article. You shall notify us in writing of the names of any such co-owners.

11. This agreement (and any dispute, proceeding, claim or controversy in relation to it) is subject to English law and the jurisdiction of the Courts of England and Wales. It may only be amended by a document signed by both of us.

Signed ______________________________________

Print name ______________________________________

Date ______________________________________