内容分析：概念、类型与方法

许汉成

（黑龙江大学俄语语言文学中心，黑龙江 哈尔滨 150080）

摘　要：本文介绍内容分析的概念、类型与方法，这是一种在西方国家十分常见的研究方法，具有巨大的应用价值。作者认为，内容分析是语言学理论的一个应用领域，另一方面，内容分析的思想和方法也向语言学家提出了许多新课题，提供了许多新方法，值得我国的语言学家关注。

关键词：内容分析；计算语言学；文本；定量分析

中图分类号：H085　　**文献标识码：**A

1 内容分析的基本概念

内容分析（контент-анализ, content analysis）是一种研究方法，在进行内容分析时，研究人员考察单个文本或者一定数量文本的集合，研究文本里某些词或者概念、范畴是否出现以及出现了多少次，分析词语或者概念、范畴之间的关系，从而就文本里的信息、作者、受话人及文本所代表的人物、文化和时代进行推断。文本可以是书、书的个别章节、小品文、采访报道、报刊标题和文章、历史文献、日记、演讲、剧本、广告词等，实际上，任何言语交际作品都可以成为文本分析的对象（Neuendorf, Kimberly　2002；Stemler, Steve 2001）。

原则上，内容分析可以采用定性分析和定量分析方法，但是，定量分析在内容分析中占主导地位。目前，内容分析基本上是借助计算机进行。定量内容分析关心反映特定文本内容的词语或者范畴的出现频数（frequency），根据频数或者频率（relative frequency）做出结论；定性分析则不同，特定词语哪怕只出现一次或者根本不出现，定性内容分析也能做出某种结论。举一个简单的例子进行说明。在 50 年代，西方分析家通过对《真理报》文章的定量分析，发现“斯大林”这个名字的出现次数迅速减少，他们由此得出一个结论，即斯大林原来的追随者在尽力与斯大林拉开距离；另一方面，定性分析专家注意到，在某位领导纪念苏联卫国战争胜利日的公开讲话里，斯大林的名字只字未提，定性分析专家可以根据这样一个简单的事实做出类似的结论。

应该注意的是，在进行内容分析时，研究者的主要兴趣并不在于各种参数、变量，而在于隐藏在参数、变量背后的现实，即文本作者的个性特征、作者想通过文本达到的目标、受话人的特征、各种社会生活事件，等等。

内容分析的应用领域可以说是令人眼花缭乱，从市场调研、大众媒体调查到文学、修辞学、民族学、文化学研究，甚至于社会学、政治学、心理学和认知科学等许多研究领域。内容分析与计算语言学、社会学和心理语言学关系尤为密切，一个内容分析系统往往是一个集人工智能大成的项目。内容分析是建立在语言学的基础上，试图从语言材料中得到客观世界的信息，是一个值得语言学家十分关注的领域。

2 内容分析的主要类型和方法

内容分析大致可以分为概念分析（conceptual analysis）和关系分析（relational analysis）。举个例子来说明这个问题。直觉告诉我，某诗人常写到饥饿，在进行概念分析时，我们就可以考察诗人诗集里“饿”、“饥饿”、“饥荒”、“忍饥挨饿的”、“饥民”这一类词的出现次数；在进行关系分析时，我们就要进一步研究概念之间的关系，研究“饥饿”、“饥荒”这组词前后常出现哪一类词、范畴或概念（Stemler, Steve 2001）。

概念分析应该从确定研究课题和选择文本样本开始。研究人员将文本分割成小的、可以操作的单位，如词、短语、句子或者主题，人工或者在软件的帮助下进行必要的标注，对文本进行编码。编码是一个选择性简化过程，它将文本压缩为范畴，每个范畴包含一个或者若干词或短语。范畴编码一般分为八个步骤：

1） 决定分析深度，即决定哪些或哪一类词或短语需要编码；
2） 决定需要编码的不同概念的数量；
3） 决定是对一个概念的存在/不存在编码，还是对概念的出现频数进行编码；
4） 决定怎样区分不同概念，例如决定词的不同形式、派生词是否归入同一概念，甚至于决定是否将隐含有某种特定意义的词也归入概念或者范畴（包括技术术语、黑话、委婉语）里，等等；
5） 建立对文本进行编码的规则，保持编码的一致性；
6） 决定怎样处理不相关信息，即决定是排除不相关信息呢，还是修改编码规则；
7） 利用手工或者计算机软件对文本进行编码；
8） 分析结果，就研究课题做出结论。

像概念分析一样，关系分析不仅仅考察文本或者文本集合里的特定概念的存在与否，还要分析概念之间的语义关系。

进行内容分析的主要理论角度有语言学和认知科学。从语言学角度进行内容分析时，研究者的注意力集中在语言单位（如词或短语），按照一定（如情感或心理）尺度给语言单位评分。从认知科学角度进行内容分析时，分析者试图建立决策图（decision maps）和心理模型（mental models），表达分析对象的观念、信念、态度与各种文本信息之间的关系。这种研究方法一般包括下面 5 个步骤：

1） 确定概念；
2） 定义关系类型；
3） 在第一步和第二步基础上对文本进行编码；
4） 对陈述进行编码；
5） 用图形显示结果，并且利用统计数字进行分析。

总而言之，内容分析（包括概念分析和关系分析）主要包括以下步骤：

1） 确定研究课题，即明确研究的问题；
2） 选择一个或者若干分析样本；
3） 确定分析类型，一般存在 3 种类型：情感提取（affect extraction）提供对文本里的显性词语的情感评分，分析说话人或者作者的情感或心理状态；接近分析（proximity analysis）关心文本里显性概念的共现（concurrence），通常是定义一个以词语为单位的长度，然后到文本里去扫描，建立概念矩阵（concept matrix），相关的、共现的概念可能会提供某种信息。在接近分析中，聚类（clustering）、分组（grouping）、标度（scaling）等方法也很有效；认知图（cognitive mapping）试图在前面两种方法基础上前进一步，将关系可视化、图形化。不论是情感提取，还是接近分析，分析都按照文本原有的顺序进行，而认知图试图建立文本整体意义的模型，将其表示为代表概念之间关系图形。认知图可以展示文本、作者或说话人以及社会团体某个时期的心理模型或者世界图景；

4） 将文本简化为范畴，对词或者短语进行编码；

5） 探索概念关系。在内容分析中，强度、符号和方向三个概念在探索概念关系中起着核心作用。关系强度是指两个或两个以上关系的联系密切程度。如果概念之间的所有关系都视为平等的，关系更容易分析、比较和用图形展示。不过，给关系赋上不同的强度值，就能更加精确地描写文本里的概念关系。关系强度能够表达“除非”、“或许”、“可能”这一类词与特定文本词语、短语或概念关联程度。符号表示概念关系是正的还是负的，例如，作为股市术语，“熊”与“股市”的关系是负的，而“牛”与“股市”的关系则是正的。有的关系是有方向的，如“X 蕴含 Y”、“X 在 Y 之后出现”以及“如果 X，那么 Y”。对这种信息进行编码有利于研究某些问题，如新信息在决策过程中的作用；

6） 对关系进行编码。概念分析和关系分析的主要区别之一是关系分析对陈述（statements）编码。陈述的逻辑概念，对应于语言学里的句子。我们知道，句子里的谓词实际上表达着语义角色的性质、或者角色之间的关系；

7） 进行统计分析；

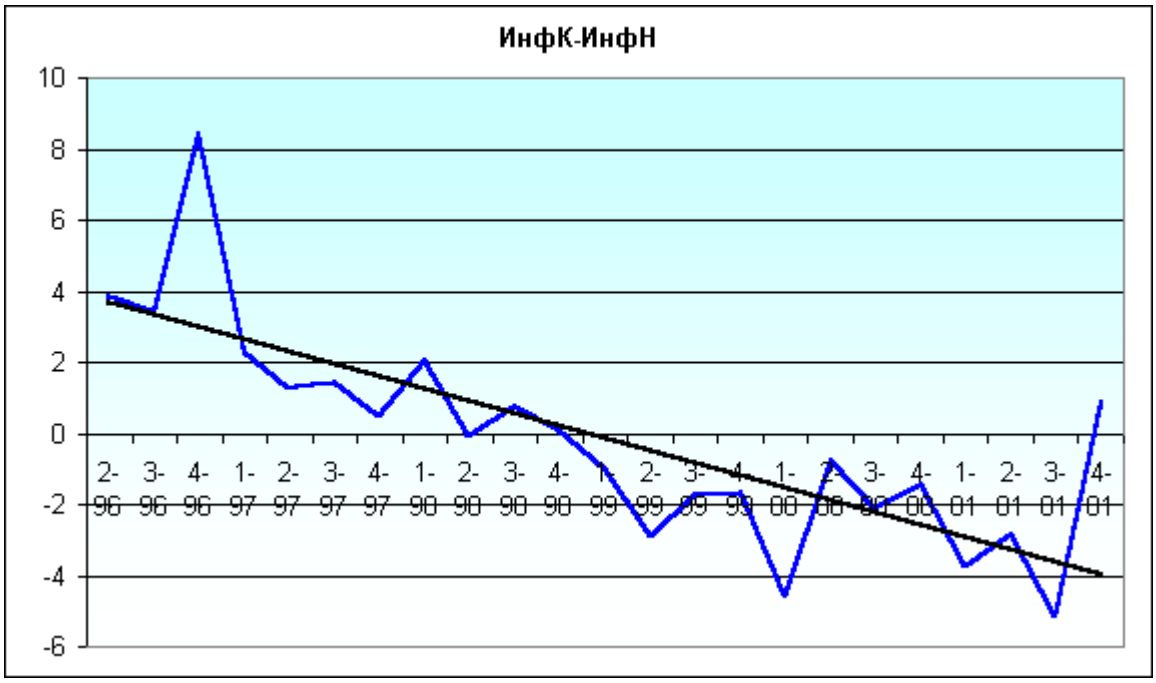
8） 将研究结果用图形表达出来。

3 内容分析的实例

不论是在英美这些发达国家，还是在俄罗斯，内容分析都有相当影响，开发了很多内容分析专家系统。ВААЛ 系统（www.vaal.ru）是俄罗斯学者沙拉克（Шалак В. 2002）开发的优秀内容分析系统。安然公司（ENRON）2001 年 12 月 2 日宣布破产，成为美国历史最大的破产案，一时轰动全球。在宣布破产前，安然公司收买审计公司，伪造了公司帐目，沙拉克利用自己的内容分析软件分析了安然公司从 1996 年 4 月到 2002 年发布的公开新闻材料，试图从这些语言材料中发现安然公司经营不佳的蛛丝马迹。

沙拉克将公司经营不善指标分为长期指标和短期指标两类，每类指标都包含了若干个具体指标。

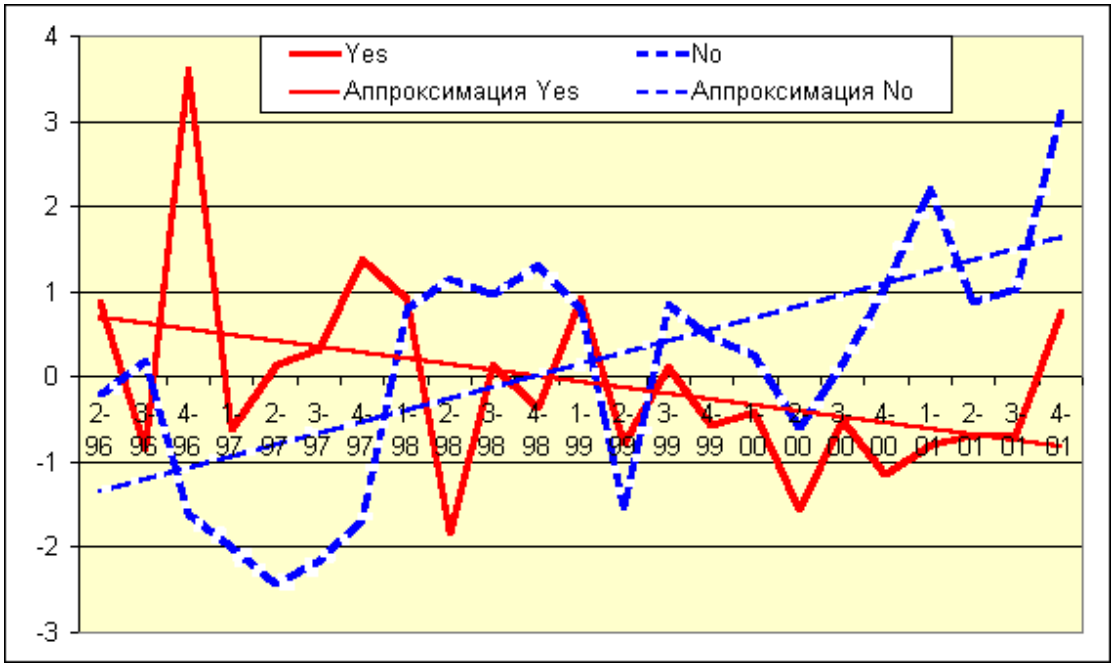
具体信息和非具体信息的比值（соотношение конкретной и неконкретной информации）是第一个公司经营不善的长期指标。这个指标分析文本内容的具体信息量（ИнфК）及文本内容的非具体信息量（ИнфН）之间的关系。安然公司的这个指标分布情况如下图所示：



*黑色直线为整个分析阶段线性逼近结果。

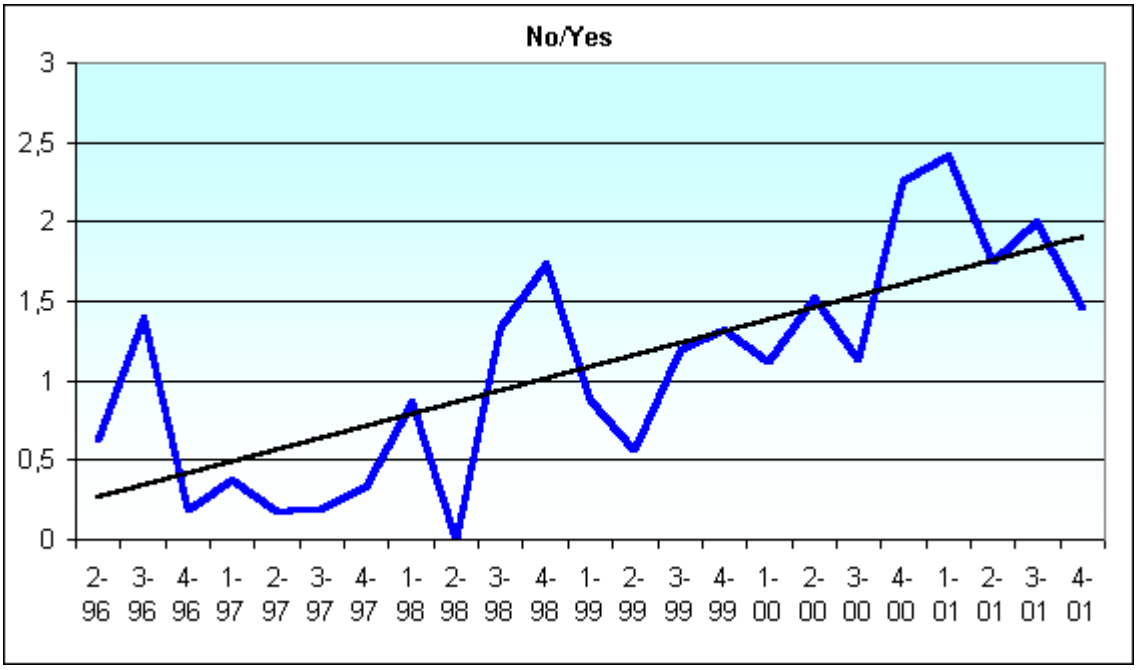
正式公开的材料里发布多少具体信息和非具体信息水平，这是公司自己的事。但这个比值的变化却不能不让人警惕，何况这种波动远远超过了±2 的偶然波动正常值范围，甚至达到±3、±4 的标准误差！具体信息和非具体信息不良趋势显示，在公布的信息中“水分”一年比一年多。该公司当然有东西需要隐瞒！

公司经营不善的第二个长期指标是新闻发布材料里同意和不同意确认比（соотношение согласия и несогласия）。人们说“不”比说“是”要困难得多，如果说“不”，那么总是应该有些理由。下图显示了研究期间“Yes”和“No”范畴的分布情况，一般认为±2 的误差



没有实质意义。

下图显示了不同意和同意确认百分比比率。



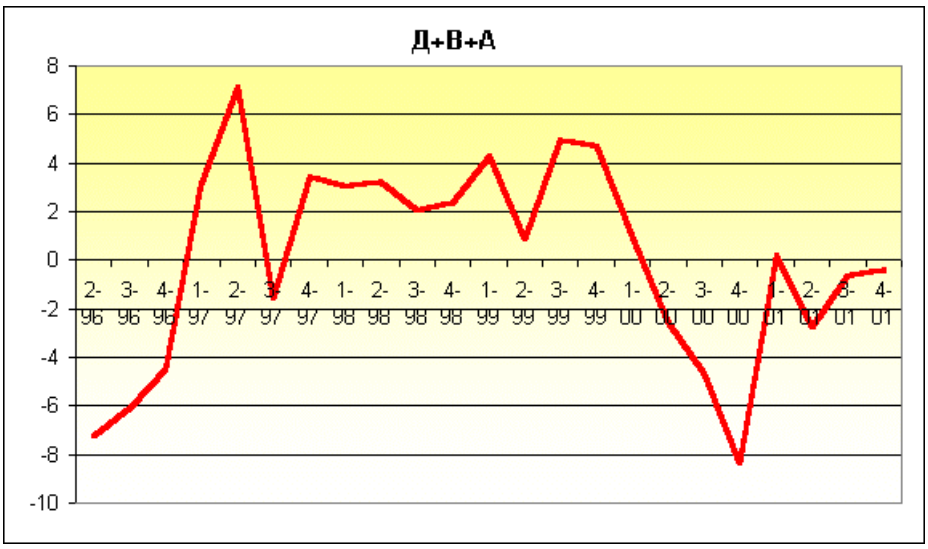
不难计算，与 1997 年相比，2001 年初这个比率增加了七倍。

对于安然公司来说，第三个长期指标为权力动机水平：

权力动机是一个心理学术语，西方学者认为这是人们的行为动机之一。对于个人而言，权力动机下降说明他心理已经满足了，而对于公司而言，说明公司安于现状、自高自大，这就孕育着不确定因素。

沙拉克还分析了 6 个短期指标：正面—负面词汇（Позитив – Негатив）、活跃性（Активный – Пассивный）、成就动机（Мотив достижения）、归属感动机（Мотив

аффиляции)、工具活动（Инструментальная деятельность）及权力+动机+归属（Д+В+А）综合指标。由于篇幅原因，这里就只给出将最后一个综合短期指标图：



可以看出，只要建立了正确的统计模型，内容分析可以从公司发布的新闻材料里发现大量潜在的、公司没有明确公布的内容。这些信息甚至比普通经济指标能传达出更多的信息。

当然，沙拉克对安然公司的研究最终是否可靠、真实，我们还需要进一步考察，研究其原始材料、统计模型、编码过程、统计图制作方法，等等。无论如何，沙拉克研究有几点值得我们重视：1）无论统计模型如何，词语、概念（或范畴）的频数、频率及其词语、概念（或范畴）的关系是内容分析的基础；2）语法学、语义学、语用学、语料库语言学、统计语言学和计算语言学的成果可以大大促进内容分析的发展，另一方面，内容分析也是理论语言学的一个重要应用领域；3）内容分析在政治、军事、经济、社会研究等各个方面具有巨大的应用价值。

4 内容分析的可靠性和有效性

任何研究方法都有可靠性（reliability）和有效性（validity）的问题（Stemler, Steve 2001）。

内容分析的可靠性是指它的稳定性（stability）、可重复性（reproducibility）以及准确性（accuracy）。稳定性是指经过一段时间编码者可以用同样的方法前后一致地对同一数据进行编码的特性。可重复性是指一组编码者可以按照同一方式对范畴成员进行分类的特性。准确性则是指文本分类符合统计标准或者规范的程度。

内容分析的有效性是指范畴是否符合结论以及能否将结论概括起来，形成普遍理论。在定义范畴时，不仅要包括显性概念，更要注意包括隐性概念，从而得到公认的范畴的定义。例如，要统计文本里“骆驼”概念范畴的出现频率，应该注意将“沙漠之舟”也包括进来。“骆驼”可以看作是显性概念，而“沙漠之舟”则是隐性变量。

在内容分析中，通过推理过程是否可以达到结论是一个富有挑战性的题目。从数据里能否推出结论，这些又是否可以用其它现象解释？例如，对于频率一类研究，一个词的第二次出现是否与第九十九次出现具有同样的权重？根据大量定量数据可以得出合理的结论，但是这些结论仍然须要进一步证明。

用计算机程序进行词频统计有时会遇到困难，影响结论的有效性。例如，英语词“mine”可以是一个物主人称代词（“我的”），也可以表示一种爆炸装置（“地雷”），还可以表示采矿用的向地下深处挖的通道（“矿井”）。在俄语里，词形 стали 可以是名词“钢铁（сталь）”的单数第二格、第三格、第四格、第六格等形式（语法歧义），也可以是动词“开始、变得

（стать）”的过去时复数形式（词汇歧义）。假如一个人发现字符“mine”出现了 50 次，其中有 17 次用于人称代词的意义，这个人实际想研究的是“mine”作为一种爆炸装置的问题，那么 50 就不是一个准确的数字。以 50 这个数字为基础做出的任何结论也都是无效的。“стали”也有类似问题。

结论的普遍性在很大程度上取决于怎样定义范畴以及范畴是否可靠，要符合上面所说的稳定性、可重复性和准确性。定义的范畴必须准确地度量所要度量的概念和/或项目，规则的建立也是如此。

5 结论

可以看出，内容分析的目标是通过对单个文本或者大量文本集合的分析，揭示文本作者、文本内容和外部世界某些特性。文本分析的关键是编码，编码将大量文字信息压缩成为概念、范畴，为统计分析奠定了基础。内容分析的主要方法是统计，或者说是定量分析。强大的计算机硬件和内容分析软件、专家系统为内容分析提供了便利，使得内容分析发展水平大幅提高。内容分析在社会政治、经济、文化分析调查、预测分析方面有巨大的应用价值。

内容分析涉及语言学、统计学、心理学、社会学、计算机软件工程等多个学科。理论语言学是内容分析的基础，与内容分析关系最密切的有：词法学（词语的形态变化）、语义学（同义、近义等关系）、句法学（配价理论）、统计语言学（统计模型）、语料库语言学和计算语言学（研究方法和软件工具）。从计算语言学的角度看，内容分析专家系统同机器翻译一样也是一个综合性、集大成的工程，包括了语言单位（词、句、段落）的切分、词法分析和句法分析、建立范畴及确定范畴词语、建立统计对象统计模型等诸多具体过程。计算语言学主要有两大流派：经验主义和理性主义（翁富良，王野翊，1998）。经验主义重视具体语料，注意建立统计模型和应用统计方法，内容分析似乎与经验主义的计算语言学更接近。但是，另一方面，内容分析建立语义网络、心理模型的工作，又让人想起计算机语言学理性主义流派建立词汇知识库的工作。内容分析建立的范畴基本上不是语言学性质的，如果我们以句子为单位建立更有语言色彩的语义—语用范畴会怎么样？我们设想，内容分析的方法能够帮助我们建立语义网络、词汇知识库。

现代语言学的一个显著特点是跨学科性，自然科学的研究方法和手段越来越深入渗透到语言研究领域，认知语言学、统计语言学、心理语言学、计算语言学、语料库语言学这些新兴领域与自然科学的关系尤为密切。语言学与自然科学的融合为语言研究提供了崭新的研究方法，拓展了语言学的应用领域，同时也向语言学研究提出了新的挑战。内容分析是语言学理论的实验田，同时也为语言学研究提供了许多新方法，提出了许多新课题，值得我国的语言学家思考和研究。

参考文献

[1]Neuendorf, Kimberly A. 2002 The content analysis guidebook. Thousand Oaks, Calif. : Sage Publications.

[2]Stemler, Steve. 2001 An overview of content analysis. Practical Assessment, Research & Evaluation, 7(17). Available online: http://ericae.net/pare/getvn.asp?v=7&n=17.

[3]Шалак В. Компьютерный контент-анализ текстов как метод экономической разведки (полный вариант статьи). Available online: http://www.vaal.ru/show.php?id=73

[4]翁富良、王野翊 1998 计算语言学导论[M]，北京：中国社会科学出版社。

Content Analysis: Concept, Types and Methodology

XU Han-cheng

(Center for Russian Language and Literature Studies of Heilongjiang University, Harbin 150080, China)

Abstract: The paper introduces the concept, kinds and methodologies of content analysis, which is well-known among western researchers, illustrates its great application potentials. The author argues that content analysis deserves close attention of linguists because it is an important application area of linguistic theories, in another hand, the idea and methods of content analysis also raised some new linguistic problems, and presented some new research methodologies for linguists.

Key words: content analysis; computational linguistics; text; quantitative approach

收稿日期: 2003-07-23

作者简介: 许汉成（1965 - ），男，陕西汉中人，副教授，主要研究方向：俄语语言学，计算语言学。

[责任编辑：靳铭吉]