语篇连贯性的量化测量*

——基于向心理论的研究

王德亮

北京师范大学外文学院 北京 100875

E-mail: bright7883@126.com

摘要：语篇连贯具有层级性，是一个模糊的、抽象的概念。为了有效地比较不同语篇的连贯性，本文基于向心理论提出了语篇连贯量化测量的具体方法，推导出了连贯度的计算公式，并选取实例进行了分析。通过量化测量，可以发现连贯性的微妙差异。量化测量的方法也可用于自然语言处理的研究之中，如作文自动评分系统，自动文摘内容的连贯性评测等。

关键词：语篇，连贯性，测量，向心理论

Quantified Measurement of the Discourse Coherence

—— A Study Based on Centering Theory

Wang Deliang

School of Foreign Languages and Literatures, Beijing Normal University, Beijing 100875

E-mail: bright7883@126.com

Abstract: The coherence in discourse is a hierarchical notion and it is vague and abstract in a certain degree. In order to compare the coherences in different discourses effectively, the present paper proposes a method capable of measuring the coherence quantitatively. A formula of measurement is postulated and examples are analyzed based on it. Through quantified measurement, fine difference in coherence can be identified. This method can be applied to natural language processing, such as automated essay scoring system, coherence measurement in automatic abstracting.

Key words: discourse, coherence, measurement, Centering Theory

1. 引言

语篇连贯性是一个层级概念，即使是连贯的语篇，也有连贯性的强弱之分。从这个意义上说，语篇连贯性是一个非常模糊的概念，如果我们对任何语篇的连贯性都能够给出量化的测量，那么语篇分析和语篇处理（discourse processing）将会变得更加清楚明了，更加容易操作。

本研究的理论基础是向心理论(centering theory)(Grosz, Joshi and Weinstein 1983, 1995; Walker, Joshi and Prince 1998)。向心理论是关于语篇局部连贯（local coherence）的理论模型。此理论对语篇连贯进行了形式化模拟，提出了一套完整的规则和制约条件，因其简洁、易操作、易处理，非常具有吸引力。自被提出之日起，向心理论就引起了多方面研究者的关注，包括语言学家、认知科学家和计算科学家。国内一些学者也进行了向心理论的相关研究（苗兴伟 2003；王德亮 2004；熊学亮、翁依琴 2005；许余龙 2008a，2008b 等等）。本文基于向心理论提出了语篇连贯量化测量的具体方法，推导出了连贯度的计算公式，并选取实例进行了分析说明。

*本文承教育部人文社会科学研究项目基金资助（项目号：08JC740001），特此致谢。

2. 向心理论的主要内容

向心理论是表现语篇语义概念凸显性（salience）运行机制的模型。其核心概念是中心（center），但它界定的中心与别的文献的界定不同。它把语篇中的所有语义实体[1]（semantic entity）都称为中心。一个语句中所有的中心都有可能成为下一句所关注的焦点，它们都被称为下指中心（forward-looking center, 简称 Cf），下指中心的集合，记作 Cf（U_i, D）[2]。Cf 中有一个特殊的成员与上文所提及的某个实体存在某种联系，被称为上指中心（backward-looking center, 简称 Cb）。Cf 根据语篇凸显性可以进行排序，排在最靠前的成员，即，最凸显的成员，被称为优选中心 Cp（preferred center），Cp 最有可能成为下一句的焦点。

向心理论的理论框架中还包括三个制约条件（Walker, Joshi and Prince 1998；王德亮 2004）。对于由语句 U_1……U_m 组成的语篇片段 D 中的每一个语句 U_i：

1）只有一个上指中心 Cb（U_i, D）。

2）下指中心集合 Cf（U_i, D）中的每一个成分都必须在 U_i 中实现（realize）。

3）上指中心 Cb（U_i, D）在 U_i 中所实现的下指中心的集合 Cf（U_{i-1}, D）中凸显性最高。

另外向心理论对于紧邻的两个语句之间的过渡状态（transition state）也进行了形式化的描述。语句 U_{i-1} 与下一个语句 U_i 之间过渡方式的划分基于两个因素：U_{i-1} 与 U_i 中的 Cb 是否相同；U_{i-1} 中的 Cb 是否与 U_i 中的 Cp 相同，即：

（1） $Cb(U_i) = Cb(U_{i-1})$, 或 $Cb(U_{i-1}) = [?]$

（2） $Cb(U_i) = Cp(U_i)$

（其中 $Cb(U_{i-1}) = [?]$ 表示 $Cb(U_{i-1})$不存在的情况，比如在语篇的开始。）

过渡状态的定义可归纳如下 (Brennan, Friedman & Pollard 1987)：

	$Cb(U_i) = Cb(U_{i-1})$, 或 $Cb(U_{i-1}) = [?]$	$Cb(U_i) \neq Cb(U_{i-1})$
$Cb(U_i) = Cp(U_i)$	延续（continue）	流畅转换（smooth-shift）
$Cb(U_i) \neq Cp(U_i)$	保持（retain）	非流畅转换（rough-shift）

根据向心理论的规则（2）（可参见 Walker, Joshi and Prince 1998；王德亮 2004），过渡状态按一定的顺序排列，延续过渡优于保持过渡，保持过渡优于流畅转换过渡，流畅转换过渡优于非流畅转换过渡。

3. 向心理论对语篇连贯性的解释

向心理论的目的之一是阐述导致语篇连贯性差异的语篇处理因素。下面我们先从一个著名的例子谈起。

(1) a Jeff helped Dick wash the car.

b He washed the windows as Dick waxed the car.

c He soaped a pane.

---

[1] 语义实体是指为语篇贡献语义概念的语句中的名词或相当于名词的结构所体现的实体。

[2] Cf（U_i, D），表示语篇片段 D 中的语句 U_i 中所有的语篇实体的集合。

(2) a Jeff helped Dick wash the car.

b He washed the windows as Dick waxed the car

c He buffed the hood. （转引自 Walker, Joshi and Prince 1998）

单纯从语篇理解的语义理论或语用推理的角度看，例(1)和例(2)中的两个语篇在连贯性上似乎并没有多大差异。(lc)中的代词 he 回指 Jeff，因为 soaped 所表达的动作是 washed 所表达的事件的一部分; (2c)中的代词 he 回指 Dick，因为 buffed 所表达的动作与 waxed 所表达的事件有关。但是，从向心理论的角度看，例(1)中的语篇比例(2)中的语篇更为连贯。根据向心理论的制约条件、规则、下指中心的凸显性、过渡状态，例(1)和例(2)可以被标记为：

(1) a. Jeff helped Dick wash the car.

Cb= [?]； **Cf** ={ JEFF, DICK, CAR }；过渡状态=NO CB

b. He washed the windows as Dick waxed the car.

Cb=[JEFF]； **Cf** ={JEFF, WINDOWS, DICK, CAR}；过渡状态=延续

c. He soaped a pane.

Cb= [JEFF]； **Cf** ={JEFF, PANE}；过渡状态=延续

(2) a. Jeff helped Dick wash the car.

Cb= [?]； **Cf** ={ JEFF, DICK, CAR }；过渡状态=NO CB

b. He washed the windows as Dick waxed the car.

Cb=[JEFF]； **Cf** ={JEFF, WINDOWS, DICK, CAR}；过渡状态=延续

c. He buffed the hood.

Cb=[DICK]； **Cf** = {DICK, HOOD}；过渡状态=流畅转换

因为例(1b)和(lc)的 Cb 一直是 Jeff，整个语篇片段中，Jeff 的凸显性最高，认知处理的难度小，消耗小。但例(2)中，(2b)的 Cb 为 Jeff，而(2c)的 Cb 则转换为 Dick。Cb 的转换则意味着认知处理的难度增大，认知负担和消耗都会增大。所以借助于向心理论，我们可以清楚明了地看出例(1)的连贯性比例(2)的连贯性强。

4. 量化测量的操作方法

向心理论为语篇定义了一系列的过渡状态(Brennan, Friedman and Pollard 1987; Grosz, Joshi, and Weinstein 1995)。根据过渡状态的排序原则[3]，我们可以做出如下理解：一个语篇中，延续过渡越多，语篇就越连贯；反之，语篇中转换过渡越多，语篇的连贯性就越弱。如果我们能够计算出不同过渡状态的数量，我们就可以得到语篇的连贯度得分。

一个语篇片段连贯度得分[4]的计算过程如下：首先给基本的过渡状态一个基本得分，即，赋予每个过渡状态一个权重（weight），如表 1 所示，

表 1　过渡状态连贯度得分表

过渡状态	权重 (Weight)
延续过渡（continue）	4
保持过渡（retain）	3

[3] 即，延续过渡 > 保持过渡 > 流畅转换过渡 > 非流畅转换过渡

[4] 亦可参见 Cristea , Ide & Romary（1998）关于平滑度得分（smoothness score）的提法。

流畅转换过渡（smooth-shift）	2
非流畅转换过渡（rough-shift）	1
无过渡（No Transition）	0

然后把整个片段中每个过渡的权重求和，其和再除以过渡的数量，这样就可以得出这个语篇片段的连贯度。可以形式化表示为：

对于由语句 U_1……U_m 组成的语篇片段 D，其连贯度 Coherence 的计算公式为：

$$Coherence = \frac{\sum_{i=1}^{m-1} w_{Transition\ U_i U_{i+1}}}{m-1}$$

通过连贯度的取值，我们可以量化出一个语篇片段的连贯程度。连贯度是一个指数，它的取值范围应该在 0 至 4 的区间上，即，如果连贯度等于 4，它表示语篇达到了最佳的连贯状态；而如果它等于 0，表示语篇最不连贯，可以说，语篇无任何连贯性。如下例：

（3）小张每天 5 点半起床，红色的封面最好看。

这两个小句之间无任何连贯性，听起来前言不搭后语。如果按照向心理论的术语来分析，前后两句的 Cb 和 Cp 之前没有任何关联，所以它们之间无过渡状态，权重为 0，连贯度为 0。

根据以上公式，我们还可以计算出例（1）的连贯度为 4，而例（2）的连贯度为 3，所以例（1）比例（2）更加连贯。如此量化测算，语篇连贯性就可以转化为容易把握的数值，利于下一步的比较和计算。

5. 实例分析

我们来看另外一组经典的例子(Grosz, Joshi and Weinstein 1995)

(4) a. John went to his favorite music store to buy a piano.

b. He had frequented the store for many years.

c. He was excited that he could finally buy a piano.

d. He arrived just as the store was closing for the day.

(5) a. John went to his favorite music store to buy a piano.

b. It was a store John had frequented for many years.

c. He was excited that he could finally buy a piano.

d. It was closing just as John arrived.

根据向心理论模型，我们可以标记出例（4）和例（5）的向心结构：

(4) a. John went to his favorite music store to buy a piano.

Cb= [?]； **Cf**= {JOHN, STORE, PIANO}；过渡状态=NO CB

b. He had frequented the store for many years.

Cb= [JOHN]； **Cf**= {JOHN, STORE, YEAR}；过渡状态=延续

c. He was excited that he could finally buy a piano.

Cb= [JOHN]； **Cf**= {JOHN, PIANO}；过渡状态=延续

d. He arrived just as the store was closing for the day.

Cb= [JOHN]； **Cf**= {JOHN, STORE, DAY}； 过渡状态=延续

(5) a. John went to his favorite music store to buy a piano.

Cb= [?]； **Cf**= {JOHN, STORE, PIANO}； 过渡状态=NO CB

b. It was a store John had frequented for many years.

Cb= [STORE]； **Cf**= {STORE, JOHN, YEAR};过渡状态=非流畅转换

c. He was excited that he could finally buy a piano.

Cb= [JOHN]； **Cf**={JOHN, PIANO}； 过渡状态=非流畅转换

d. It was closing just as John arrived.

Cb= [STORE]； **Cf**= {STORE, JOHN}； 过渡状态=非流畅转换

然后再根据以上提出的连贯度计算公式，我们可以得出，例（4）的连贯度取值为 4，而例（5）的连贯度取值为 1。这两个取值差距很大，说明这两个语篇的连贯性有很大差异。这也是为什么当我们读完这两个语篇片断后，明显感觉例（4）非常连贯，全文都是围绕“John”这个中心展开。而例（5）不断变化中心，连贯性较差。

把语篇的连贯性量化之后，我们可以精确计算出连贯性的微妙差异。这样可以帮助我们解释在连贯性上存在微妙差异的语篇，因为有些语篇连贯性差异很小，仅凭语感，很难说清，而且语感因人而异，有主观性，有时不可靠。下面我们来看一组汉语的例子：

(6) 我家花园里有棵树，这种树叶子很小，呈卵圆形，冬天也不凋落。

(7) 我家花园里有棵树，这种树叶子很小，根系发达，主要生长在南方。

下面，我们可以标记出例（6）和例（7）的向心结构，为了方便理解，同时标记出回指关系（anaphoric relation）。

(6) a. 我家花园里有棵树$_j$，

Cb= [？]; **Cf**={家，花园，树};过渡状态=NO CB

b. 这种树$_j$叶子$_j$很小，

Cb= [树]; **Cf**={树，叶子};过渡状态=延续

c. ϕ_j呈卵圆形，

Cb= [叶子]； **Cf**={叶子，卵圆形}； 过渡状态=流畅转换

d. ϕ_j冬天也不凋落。

Cb= [叶子]； **Cf**={叶子，冬天}； 过渡状态=延续

(7) a. 我家花园里有棵树$_j$，

Cb= [？]; **Cf**={家,花园，树};过渡状态=NO CB

b. 这种树 i 叶子 j 很小，

Cb= [树]； **Cf**= {树，叶子};过渡状态=延续

c. ϕ_i根系发达，

Cb= [树];**Cf**= {树,根系};过渡状态=延续

d. ϕ_i主要生长在南方。

Cb= [树]; **Cf**= {树,南方};过渡状态=延续

根据连贯度计算公式，我们可以得出，例（6）的连贯度取值为 3.33，而例（7）的连贯度取值为 4，例（7）比例（6）更加连贯。但是，两个语篇片断的连贯度很接近，如果不用量化测量的话，仅凭语感，很难讲清其连贯性的差异。从这一点上，我们可以看出量化测量的优势。

6. 结语

向心理论所规定的规则和制约条件是语篇组织的倾向性准则，并不是不可违反的法则，即，如果遵守了这些准则，语篇就是理想中的连贯状态，更容易被理解和处理；如果语篇违反了某项准则，它也不会不合语法，只是它的连贯性减弱，不容易被理解和处理，认知负担会增加。基于向心理论提出的关于过渡状态的理念，我们推导出一套语篇连贯度的计算公式。根据计算公式，我们可以量化测量语篇片断的连贯性，从而发现语篇连贯性的微妙差异。

语篇连贯性的量化测量可应用于开发和改善作文自动评分系统[5]，国外一些学者已经对此有所关注，如 Karamanis (2001)和 Miltsakaki & Kukich (2000)。另外，语篇连贯性的量化测量还可以应用于自动文摘内容的连贯性评测。

参 考 文 献

[1] Brennan, S., M. Friedman & C. Pollard, 1987. A Centering Approach to Pronouns [A]. In *Proceedings of the 25th Annual Meeting of the Association for Computational Linguistics* [P], Stanford, California.

[2] Cristea, D., N. Ide & L. Romary. 1998. Veins theory: a model of global discourse cohesion and coherence [A]. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th Int. Conf. on Computational Linguistics* [P], pp.281-285. San Francisco, California.

[3] Grosz, B. J., A. K. Joshi & S. Weinstein. 1995. Centering: a framework for modeling the local coherence of discourse [J]. *Computational Linguistics,* 21(2): 203-225.

[4] Grosz, B. J., A. K. Joshi & Scott Weinstein. 1983. Providing a unified account of definite noun phrases in discourse [A]. In *Proceedings of the 21st Annual Meeting of the Association of Computational Linguistics* [P], pp. 44–50. Cambridge, Mass.

[5] Karamanis, N. 2001. Exploring Entity-based Coherence [A]. *Proceedings of CLUK4* [P]. University of Sheffield, January 2001, 18-26.

[6] Miltsakaki, E. and K. Kukich, 2000. The Role of Centering Theory's Rough-Shift in the Teaching and Evaluation of Writing Skills [A]. *Proceedings of the 38th ACL* [P], pp. 408–415.

[7] Walker, M. A., A. K. Joshi & E. F. Prince. 1998. Centering in naturally-occurring discourse: an overview [A]. In M.A. Walker, A.K. Joshi & E.F. Prince (eds.), *Centering Theory in Discourse* [C], pp.1-28. New York: Oxford University Press.

[8] 梁茂成，文秋芳，2007，国外作文自动评分系统评述及启示[J]，《外语电化教学》(5):18-24。

[9] 苗兴伟，2003，语篇向心理论述评[J]，《当代语言学》(2): 149-157。

[10] 王德亮，2004，汉语零形回指解析——基于向心理论的研究[J]，《现代外语》(4):350-359。

[11] 熊学亮、翁依琴，2005，回指的优选解析[J]，《外语教学与研究》(6):432-438。

[12] 许余龙，2008a，“语句”与“代词”的设定对指代消解的影响———项向心理论参数化实证研究[J]，《现代外语》(2): 111-120。

[13] 许余龙，2008b，向心理论的参数化研究 [J]，《当代语言学》(3):225-236。

[5] 作文自动评分是一项使用计算机进行作文评分的新技术；评分过程中计算机作为评分员自主评分，不需要人为干预。关于评分系统的介绍，可参见梁茂成和文秋芳（2007）。