doi:10.13582/j.cnki.1672-7835.2020.01.021

英汉多义词模体的计量特征

杨江[1],赵晗冰[2]

(1.湖南科技大学 外国语学院;2.湖南科技大学 人文学院,湖南 湘潭 411201)

摘　要:词的多义性是人类语言基本且普遍的特征。基于语言模体框架对英汉多义词模体进行基本的数据统计,探索两种语言多义词模体的计量特征,讨论英汉语中与多义词相关的语言个性特征,表明,英汉多义词模体符合齐普夫—曼德布洛特分布,英汉多义词高阶长度模体遵从混合负二项分布;语言模体框架下的多义词对比分析能揭示英汉语的诸多个性特征;英语多义词在意义上比汉语多义词具有更高的灵活度,在理解上具有更大的上下文依赖度。

关键词:多义词模体;计量特征;分布;长度;英汉语

中图分类号:H14　　**文献标志码**:A　　**文章编号**:1672-7835(2020)01-0152-08

本文所用的数据资料有两类。第一类是传统意义上的词典,其用途主要是从中抽取出词及其义项数目,从而形成英汉义项词典。其中,使用的英语词典为《美国传统词典(第四版)》(*The American Heritage Dictionary*),共计 75 685 个词条;使用的汉语词典为"现汉",包含 68 764 个词条。第二类是文本语料库,其用途是从真实文本中构建对应的英汉语多义词模体。英语语料是英语国家语料库(The British National Corpus)的一个样本,共计2 787 208个词,汉语语料是人民日报(1998年)标注语料库的一个样本,共计 2 531 511 个词。所有语料均以篇章为单位抽取入库,英语语料使用 Stanford POS Tagger 进行词性标注处理,汉语语料已经包含相关信息,无需另行标注。

一　语言中的模体及其构建

模体是计量语言学中的概念。计量语言学是以真实语料为基础、用精确的方法来研究语言结构与发展规律的语言学分支学科①②,它主要关注单位(unit)、属性(property)以及二者之间的关系(relation),采用实证研究方法来揭示语言现象背后的规律。自 1935 年美国语言学家齐普夫(G. K. Zipf)有关语言统计的著作出版以来,计量语言学历经 80 多年的发展,为现代语言学的科学化做出了重要贡献。计量语言学有两种互补的研究路径:聚量分析(analysis of language in the mass)和序列分析(analysis of language in the line)③,前者把语言单位看作统计上独立的无序变量,后者则认为语言单位的序列是一个与之紧密相关的特征④。长期以来,由于聚量分析的研究路径占据着统治地位,大多数计量语言学研究通过运用数学方法和数学模型来反映语言中的"群体现象",忽略了语言单位和结构在线性序列上的行为和特征。直到近些年,序列分析的研究路径才逐渐开始受到重视。在此过程中,语言模体扮演着重要的作用。

语言模体是德国计量语言学家莱茵哈德·科勒(Reinhard Köhler)于 2006 年提出的一个用以表示语言序列行为和特征的概念,最初被称作"片段"(segment),后又改称为"序列"(sequence),最

收稿日期:2019-05-22

基金项目:湖南省哲学社会科学基金项目(14YBA153)

作者简介:杨江(1978—),男,湖南湘乡人,博士,副教授,主要从事计算语言学与计量语言学研究。

①刘海涛,黄伟:《计量语言学的现状、理论与方法》,《浙江大学学报(人文社会科学版)》2012 年第 2 期。

②冯志伟:《用计量方法研究语言》,《外语教学与研究》2012 年第 2 期。

③Herdan G. *The Advanced Theory of Language as Choice and Chance*. Berlin: Springer-Verlag, 1966, p. 423.

④Pawlowski A. "Language in the Line vs. Language in the Mass: On the Efficiency of Sequential Modelling in the Analysis of Rhythm", *Journal of Quantitative Linguistics*, 1999(1):70-77.

后定名为“模体”(motif)并使用至今①②③。科勒关于语言模体的灵感来自音乐领域对乐谱中音符时值的频数进行研究时所定义的“F-motiv”概念④,它与文学领域表示文学作品中一种反复出现的因素即“母题”(motif)并不相干,也与生物学中表示蛋白质或 DNA 的一个特征序列的“基序”(motif)没有渊源。语言模体的提出,为计量语言学从横向组合关系的角度研究和分析语言问题开辟了一条崭新的道路。模体不仅是计量语言学中一个新兴的“语言单位”,而且为计量语言学研究引入了新的研究思路,带来了新的研究方法,拓展了新的研究领域,由此也引起了较为广泛的关注。

语言模体是表征一类语言单位特定计量属性的最长的连续等同或递增数值序列⑤。这是语言模体的一般性定义,为了便于理解,有必要对其再稍作说明。具体而言,这里的“语言单位”通常指音素、音节、词、短语、句子等能在时间或空间上依线性顺序展开的语音和语法单位,“计量属性”则多指频次、长度、距离、位置、组分等依附于某一语言单位之上的可度量的性质,如音素频次、词长(词所包含的音节数量)、句法距离(句法树中子节点到父节点的距离)等。由此,基于不同的计量属性可以形成不同类型的语言模体。例如,“频次模体”(F-motif)是语言单位频次的最长连续等同或递增数值序列,“长度模体”(L-motif)是语言单位长度的最长连续等同或递增数值序列。此外,“多义模体”(P-motif)涉及的是语素或词的义项值,而“多文度模体”(T-motif)考察的是语素、词或者句法结构的多文度(polytextuality)。定义中的“数值序列”是依据给定的语言单位属性获得的一串数字,是数值化的属性信息,一般用圆括号分隔。这串数字的特点是按非递减的方式排列,并保留了原有语言单位之间的分隔和顺序关系,因而其长度既是序列中数字的个数,又等于原始语言单位的数目。数值序列中数字的个数即为模体的长度。我们称由两个或两个以上模体构成的分组模体序列为模体组。下面以两个句子实例阐释语言模体的构建过程。

(1)Success comes to those who dedicate everything to their passion in life.

(2)农业是国民经济的基础,8 亿农民的富足、稳定至关重要。

表 1 是英语例句的词长模体组,词的长度按音节计数,据此,句中每一个词的长度属性都获得了一个对应的数值。换句话说,词对应的数值表示了该词的音节长度信息,整句(表 1 第 1 行)由此转换为一个数字序列(表 1 第 2 行)。比较相邻数字的大小,在所有递减处进行分割,用分隔符表示分割的起止位置,依次得到“(2)、(1-1-1-1-3-3)、(1-1-2)、(1-1)”共计四个模体,该句的词长模体组则为“(2)(1-1-1-1-3-3)(1-1-2)(1-1)”(表 1 第 3 行)。表 2 是汉语例句的

表 1　例句(1)的词长模体组

Success	comes	to	those	who	dedicate	everything	to	their	passion	in	life	.
2	1	1	1	1	3	3	1	1	2	1	1	
(2)	(1-1-1-1-3-3)						(1-1-2)			(1-1)		

表 2　例句(2)的多义模体组

农业	是	国民经济	的	基础	,	8 亿	农民	的	富足	、	稳定	至关重要	。
1	11	1	6	2		1	1	6	1		3	1	
(1-11)		(1-6)		(2)		(1-1-6)			(1-3)			(1)	

①Köhler R. “The Frequency Distribution of the Lengths of Length Sequences”, *Favete Linguis*: *Studies in Honour of Viktor Krupa*. Bratislava: Slovak Academic Press, 2006, pp. 145-152.

②Köhler R, Naumann S.“Quantitative Text Analysis Using L-, F- and T-Segments”, *Data Analysis, Machine Learning and Applications*: *Studies in Classification, Data Analysis, and Knowledge Organization*. Berlin: Springer, 2008, pp. 637-645.

③Köhler R. “Linguistic Motifs”, *Sequences in Language and Text*. Berlin: de Gruyter, 2015, pp. 89-108.

④Boroda M. “Häufigkeitsstrukturen musikalischer Texte”, *Sprache, Text, Kunst*: *Quantitative Analysen*. Bochum: Brockmeyer, 1982, pp. 231-262.

⑤Köhler R. “Linguistic Motifs”, *Sequences in Language and Text*. Berlin: de Gruyter, 2015, pp. 89-108.

词语多义模体组,由六个不同模体构成,各词语的义项值依据“现汉”给出,所有标点均忽略不计,最后按定义形成分组的、属性值非递减的模体以及整句的模体组。

模体可以基于传统的语言单位来构建,小到音素,大到句子甚至超句单位都可以形成模体。基于传统语言单位定义的某个类型的模体,通过真实语料的生成,转换成一个个新的语言单位,可以在计量语言学的理论框架下,将其按照传统语言单位的方式来观测、处理和分析。同时,语言模体可以为任何语言属性所定义,不论是易于计量的属性还是难以计量的属性都能得到较好、有效、灵活的处理。从选定语言单位和计量属性,到定义符合预定研究目标的语言模体,再到通过话语或文本构建实际模体的过程,称为模体的操作化(operationalisation)。

作为一个新的“语言单位”,模体具有四个优势特征:

(1)分割有穷性。模体是一串数字序列,同质的模体随着话语或文本的延展从第一个原始语言单位开始构建,到最后一个单位结束。由此,任何话语或文本都能被穷尽地分割为若干个模体。模体的生成一般以具有完整意义的话语和文本为单位。

(2)分割无歧性。模体获取的是语言单位的一类计量属性,用数值进行表征,避免了模糊性。一个语言单位的后继在数值上要么大于等于当前单位,要么小于当前单位。符合前一种情形的,该后继包含在当前模体中;符合后一种情形的,则它是一个新模体的起始。话语和文本的第一个单位没有前驱,是第一个模体的开始;最后一个单位没有后继,它既可能是最后一个模体的结束,又可能是长度为 1 的最后一个模体。

(3)尺度可变性。模体尺度的可变性在于其定义是能够重复使用的,因而可以在已有模体的基础上定义新的模体。例如,我们可以在例句(1)的词长模体上继续定义新的长度模体,形成如下模体组:“(1-6)、(3)、(2)”,其中的“1、6、3、2”分别为原有四个词长模体的长度;还可以在例句(2)的多义词模体上定义新的频次模体,即为原有六个不同模体赋予频次数值后再形成模体。这种形式的模体称为“高阶模体(higher order motifs)”。高阶模体以高阶的长度模体和高阶的频次模体最为常见,其第一阶构筑在语言单位的属性上,其余阶则基于长度或频次而形成。模体的阶数和模体的数量成反比,即模体的阶数越高,得到的模体数量就越少。

(4)分布规律化。与其他传统意义上语言单位类似,模体的频数与其频数秩(序号)的分布符合齐普夫—曼德布洛特(Zipf-Mandelbrot)定律。分布定律描述了语言结构在语言系统和语言使用中的定量特征,特定分布定律的不同参数则体现了不同语言结构和属性的差异。

在科勒关于模体的定义中,句子边界被打破,模体可以跨越两个或两个以上的句子。不论一个篇章包含多少个句子,理论上都可能只被表示为一个多义词模体组:起始于篇章的第一个词,结束于篇章的最后一个词,所有的标点符号都被忽略。与这种处理方式不同,本文为模体增加了两个约束条件:(1)任何模体必须在一个句子结尾处停止;(2)文本标点是模体序列的构成成分。理由有二:其一,科勒关于语言模体的灵感来自于 Boroda 对乐谱中音符持续时长的描写①,我们认为,语言毕竟与音乐不同,音符序列是连续的,而话语有自然停顿;其二,将话语中的停顿如实地在模体序列中反映出来,更符合语言的真实情况。例(1)、(2)的多义词模体序列已经包含了上述约束条件。

我们编制了程序从文本语料库中分别构建英汉多义词模体。表 3 是两种语言多义词模体的一些基本信息。

表 3　英汉语多义词模体基本信息

语言	词数	句子序列	模体例数	模体型数	模体平均长度	义项均值
英语	2 787 208	129 341	1 414 757	72 734	1.97	4.759 9
汉语	2 531 511	101 095	989 440	9 363	2.558 5	2.440 4

由表 3 可知,尽管所使用的英汉语语料规模在词和句子数量上基本相当,但两种语言所产生的多义词模体数量具有较大的不同:英语多义词模体例数较汉语更多。这和模体的平均长度相

①Köhler R. “Linguistic Motifs”, *Sequences in Language and Text*. Berlin: de Gruyter, 2015, pp. 89-108.

关,即模体平均长度与例数成反比:给定相同词数的语料,模体平均长度越大,则例数越少。表中两种语言模体平均长度的差异性也较显著。

表 3 中两种语言表现的最显著差异是模体型数(绝对数量)。一种更为科学的比较方式是不仅考虑模体型数,还同时参考对应的例数,这里我们借用语料库语言学中常用的型例比(type/token ratio)来考察这一差异。通过计算,英汉语多义词模体的型例比分别为 5.141 1%和 0.946 3%,差异非常显著。模体型例比反映了语言中模体的丰富度:比值越高,该语言中的模体就越丰富,意味着多义词在组合关系上具有更大的变化性。可以说,英语多义词的组合变化远远高于汉语。

词的义项均值本身和多义词模体并无直接、必然的联系,然其与语言的属性有关。考虑到后文的讨论要涉及两种语言特征的比较,因而将其呈现在此。数据显示,英语词的义项均值几近汉语的两倍。与英汉义项词典中词的义项均值(分别为 3.470 1 和 1.348 9)相比,二者的数值均有了较明显的增长,说明真实文本中多义词的使用有一定的活跃度。另一方面,由于英语语料中词的义项均值与单义词的义项值 1 相比有较显著的差异,因此英语多义词的使用占据绝对优势,而汉语语料中该数值与 1 相比差异不太显著但又具有一定的区分度,因此汉语多义词和单义词的数量大致相当。此外,从汉语词的义项均值 2.44 与汉语多义词占 49%的关系推测,义项均值 2.5 或许是一种语言多义词与单义词占比平衡的“支点”。

二　英汉多义词模体的分布特征

传统的语言单位都遵循特定的统计分布,最著名的是齐普夫定律:文本中词的频数(frequency)与其频数秩(即排列序号,rank)之间具有反比例关系。语言模体作为一种具有适宜计量、聚焦组合等优势的“语言单位”,其分布特征也可能存在普遍性的规律。科勒在研究模体分布定律时,提出了一个理论假设:语言模体在分布上同传统的语言单位相似,均遵从齐普夫—曼德尔布罗特定律。他对意大利语文本中词的音节长度模体进行了验证,实验结果支持上述假设①。本节探讨多义词模体的分布特征,以进一步验证科勒的假设。

我们使用 Altmann Fitter 软件(3.1 版)②进行统计分析,该软件内置的 200 余种分布函数能自动拟合输入的各类数据。由于从英语语料库中获取的多义词模体数量太多(详见表 3),超过了 Altmann Fitter 所能处理的上限,我们从该库中进行二次抽样,得到一个由 362 745 个词组成的较小样本语料库,重新组建英语多义词模体。本节实验所用的实际数据如表 4 和表 5 所示。

表 4　二次抽样后的英语多义词模体的 rank-frequency 分布(排名前 20 位)

序号	多义词模体	频次	序号	多义词模体	频次
1	(1)	12 756	11	(10)	2 098
2	(22)	8 427	12	(16)	1 986
3	(1-1)	4 066	13	(4)	1 968
4	(21)	3 105	14	(30)	1 927
5	(8)	2 893	15	(26)	1 872
6	(29)	2 552	16	(9)	1 802
7	(6)	2 210	17	(7)	1 600
8	(11)	2 205	18	(19)	1 478
9	(12)	2 173	19	(14)	1 466
10	(15)	2 136	20	(3)	1 432

表 5　汉语多义词模体的 rank-frequency 分布(排名前 20 位)

序号	多义词模体	频次	序号	多义词模体	频次
1	(2)	67 594	11	(5)	17 720
2	(1)	46 540	12	(1-1-2)	16 619
3	(1-10)	41 942	13	(1-4)	15 710
4	(3)	36 900	14	(6)	14 023
5	(1-2)	36 587	15	(1-6)	13 689
6	(1-1)	29 375	16	(1-5)	13 648
7	(4)	24 564	17	(8)	12 962
8	(1-3)	22 769	18	(1-1-1)	11 259
9	(1-8)	20 572	19	(2-10)	10 897
10	(1-1-10)	18 697	20	(1-14)	10 837

AltmannFitter 输出的结果表明(根据 C 和 R^2 值),英汉语多义词模体的 rank-frequency 数据均符合齐普夫—曼德尔布罗特分布。数据拟合结果的详细信息如下:

(1)英语

参数估计:a =1.085 2,b =1.195 5,n =1995 4,

①Köhler R. “Linguistic Motifs”,*Sequences in Language and Text*. Berlin: de Gruyter,2015, pp. 89-108.

②https://www.ram-verlag.eu/software-neu/software

统计检验结果:X^2 = 7 733.447 2,$P(X^2)$ = 0.000 0,DF = 16 633,C = 0.042,R^2 = 0.958 6;

(2)汉语

参数估计:a = 1.577 5,b = 7.952 4,n = 9 363,统计检验结果:X^2 = 11 792.973,$P(X^2)$ = 0.000 0,DF = 9 359,C = 0.011 9,R^2 = 0.991 7。

上述结果中字母的含义为:a,b 是齐普夫—曼德尔布罗特分布的参数,n 是模体型数。X^2 代表卡方,$P(X^2)$ 代表卡方概率,DF 表示自由度,C 是差异系数,$C = X^2/N$,其中 N 为模体例数;R^2 为决定系数。

图 1 和图 2 是两种语言数据拟合的双对数坐标图。

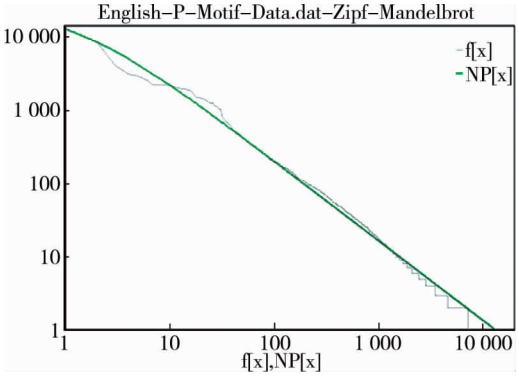

图 1　齐普夫—曼德尔布罗特分布拟合英语多义词模体的双对数坐标图

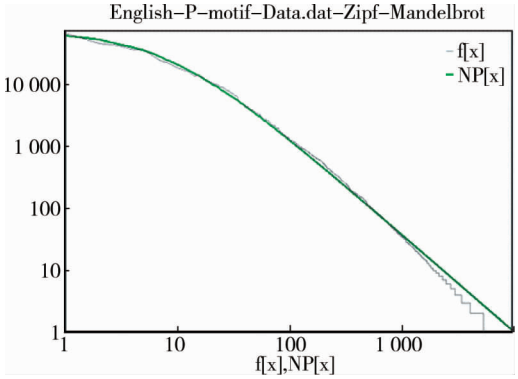

图 2　齐普夫—曼德尔布罗特分布拟合汉语多义词模体的双对数坐标图

需要说明的是,由于英汉语多义词数据集都是大样本,导致卡方检验(chi-square test)的失效。拒绝卡方检验的原因出于以下两个事实:(1)卡方检验值随着样本数据量的增长而呈线性增长;(2)语言研究中的样本数据往往是非常大的①。因此,上述结果数据中的 X^2 以及 $P(X^2)$ 的值在这里没有统计价值,评价拟合优度(goodness-of-fit)的指标应采用差异系数 C(coefficient of discrepancy)。在现代语言学研究中,差异系数具有广泛的接受度,有不依赖于自由度的优点。通常而言,差异系数越小,拟合优度则越好;当 C< 0.02 时,可以认为是一个良好的拟合,而当 C< 0.01时,则是一个非常好的拟合②。此外,在假设检验中,决定系数 R^2(coefficient of determination)也是一个可供参考的指标。决定系数用来表示数据与统计模型在多大程度上是一个好的拟合,但参考这一指标的前提是差异系数在拟合优度上具有良好的表现。据此,观察上述检验结果数据可知,对于汉语多义词的 rank-frequency 分布而言,根据 C 和 R^2 统计量,我们判断这是一个良好的拟合结果。对于英语而言,其拟合的结果可以接受,因为其差异系数(C = 0.042)与 0.02 相距很近,差异较小,其决定系数(R^2 = 0.958 6)则表明这是一个较优的拟合。综合上述情况,我们认为英语多义词模体的 rank-frequency 数据符合齐普夫—曼德尔布罗特分布,观察图 1 的拟合结果对数坐标图也能获得直觉上的认识,证明了上述判断的合理性。

与此前在其他类型的语言模体研究中获得的结论相同,英汉多义词模体在分布上具有普遍性规律,均遵循齐普夫—曼德尔布罗特分布。我们从多义词模体的角度论证了科勒提出的假设,完善了语言模体在词的多义性上的分布特征,其更深远的意义,或许在于至少从一个方面论证了多义词模体的语言单位性。在此基础上,我们才有可能将多义词模体运用到与多义词相关的传统语言学研究领域中,或开展并行研究以相互对比和验证,或从事独立研究以探索新的语言规律。本文以下的内容即分别从这两方面展开,并试图展现多义词模体的潜在应用价值。

①Antić G, Grzybek P, Stadlober E. "Mathematical Apects and Modifications of Fuck's Generalized Possion Distribution (GPD)", *Quantitative Linguistics: An International Handbook*. Berlin: de Gruyter, 2005, pp. 158-180.

②Antić G, Grzybek P, Stadlober E. "Mathematical Apects and Modifications of Fuck's Generalized Possion Distribution (GPD)", *Quantitative Linguistics: An International Handbook*. Berlin: de Gruyter, 2005, pp. 158-180.

三　英汉语多义词模体的长度

语言模体在粒度(granularity)上具有可伸缩性,这也是其与传统语言单位相比具有的一大优势。换言之,语言模体的同一定义可以重复使用,在某个语言模体的基础上构建新的其他类型的模体①。在多义词模体已经确定的前提下,其模体的长度也就相应的确定了,可以基于英汉多义词模体构造多义词的高阶长度模体。本节所使用的原始英汉语料资源如上表 3 所示。表 6 列出了使用频次最高的 20 个英语多义词模体。

表 6　英语多义词模体的 rank-frequency 分布(排名前 20 位)

序号	多义词模体	频次	序号	多义词模体	频次
1	(1)	92 403	11	(10)	16 286
2	(22)	75 201	12	(26)	15 924
3	(1-1)	26 908	13	(9)	15 055
4	(21)	22 641	14	(16)	14 187
5	(8)	21 388	15	(4)	14 101
6	(11)	19 814	16	(30)	13 636
7	(29)	18 825	17	(13)	11 748
8	(6)	17 268	18	(7)	11 565
9	(15)	17 054	19	(1-22)	11 256
10	(12)	16 334	20	(3)	10 754

拟合结果表明(根据 C 和 R^2 值),英汉语多义词模体的长度模体分布均符合混合负二项分布(mixed negative binomial distribution)。拟合结果的详细数据见表 7,两种语言的分布拟合双对数坐标图分别见图 3 和图 4。

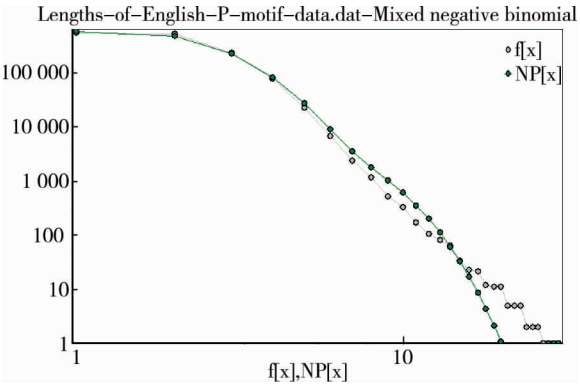

图 3　混合负二项分布拟合英语多义词的高阶长度模体双对数坐标图

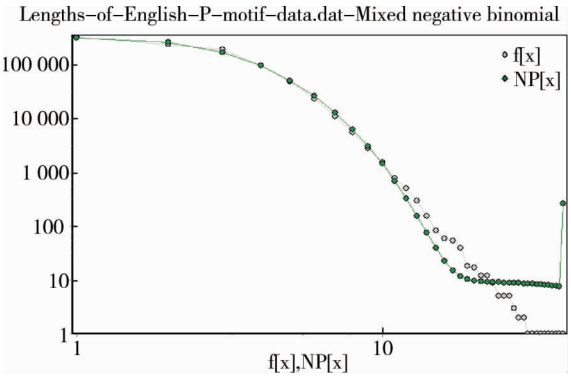

图 4　混合负二项分布拟合汉语多义词的高阶长度模体双对数坐标图

表 7　混合负二项分布拟合英汉语多义词模体的长度模体分布结果数据

语言	k	p_1	p_2	a	X^2	DF	C	R^2
英语	6.928 8	0.627	0.881 3	0.016	9 996.109 8	15	0.007 1	0.994 3
汉语	1.961 6	0.572 3	0.043 2	0.999 4	8 552.389 2	34	0.008 6	0.992 4

表 7 中,k,p_1,p_2,a 是混合负二项分布的参数,X^2,DF,C,R^2 是统计检验的结果。同前所述,判断拟合优度的指标是差异系数 C 和决定系数 R^2。

随机变量的混合分布是两种变化过程共同作用的结果②。在语言学上,对这两种过程的一个合理的解释是将其看作不同的多样性过程(diversification process)③④。就此处讨论的多义词模体及其高阶长度模体而言,这两种多样性过程均与多义词有关,其中一种过程可能表现为多义词模体中多义词的多样性,另一种过程可能是多义词的高阶长度模体的多样性所产生的结果。两种过程共同作用,导致负二项分布的形成,但每种过程在产生既定分布时所带参数有所不同,如表 7 中英汉语的 p_1 和 p_2。为了证明混合负二项分布是

①Köhler R. "Linguistic Motifs", *Sequences in Language and Text*. Berlin: de Gruyter, 2015, pp. 89-108.

②Andrei B, Köhler R, Naumann, S. "Quantitative Properties of Argumentation Motifs", *Methods and Applications of Quantitative Linguistics*. Belgrade: Academic Mind, 2013, pp. 33-43.

③Altmann G. "Modeling Diversification Phenomena in Language", *Diversification Processes in Language: Grammar*. Hagen: Rottmann, 1991, pp. 33-46.

④Altmann G. "Diversification Processes", *Quantitative Linguistics: An International Handbook*. Berlin: de Gruyter, 2005, pp. 646-658.

多义词高阶长度模体的一种合理的数学分布模型,我们还需验证多义词模体中多义词的分布符合负二项分布这一假设。下面进行验证。

上述假设中论及的"多义词的分布",是就多义词的义项类型而言的,准确的说,是词的义项值(类型)在一个模体中出现的频次情况,而且不排除单义词。比如,在多义词模体"(4-5-5-6-15)"中,第一个词的义项值为 4,则该词属于义项类型 4,该类型在模体中出现 1 次,则其频次为 1;同理,义项类型 6 和 15 的频次分别为 1;义项类型 5 出现了 2 次,其频次为 2。其他所有义项类型均没有在此例中出现,则他们的频次记为 0。

我们对所有出现在语料库中的词的义项类型进行了测试,结果显示,除义项类型 1(即单义词和标点符号的义项值)外,所有的义项类型均很好地服从负二项分布。表 8 列出了英汉语多义词模体中义项类型 3 的分布情况。

表 8　英汉多义词模体中义项类型 3 的分布数据

英语		汉语	
频次类型	频次	频次类型	频次
0	1 359 801	0	850 800
1	54 075	1	131 050
2	864	2	7 211
3	15	3	353
4	2	4	16

表 9 和表 10 是两种语言的部分义项类型分布拟合结果数据。

表 9　负二项分布拟合英语多义词模体中义项类型 2-6 的分布结果数据

义项类型	X^2	C	DF	R^2	P	N
2	3.09	0.000 0	1	1.000 0	0.999 7	1 414 757
3	42.98	0.000 0	1	1.000 0	0.999 6	1 414 757
4	173.61	0.000 1	1	1.000 0	0.999 6	1 414 757
5	0.35	0.000 0	1	1.000 0	0.994 1	1 414 757
6	541.14	0.000 4	2	1.000 0	0.999 7	1 414 757

表 10　负二项分布拟合汉语多义词模体中义项类型 2-6 的分布结果数据

义项类型	X^2	C	DF	R^2	P	N
2	1 883.58	0.001 9	4	0.999 6	0.999 2	989 440
3	704.24	0.000 7	2	1.000 0	0.999 6	989 440
4	171.17	0.000 2	2	1.000 0	0.999 4	989 440
5	6.79	0.000 0	2	1.000 0	0.974 8	989 440
6	59.65	0.000 1	2	1.000 0	0.999 6	989 440

对义项类型 1 的拟合结果表明,其服从其他类型的分布,但不符合负二项分布。造成这个例外情形的原因目前尚不可知,也许客观事实即是如此。这里我们将其视为一种特使情况,鉴于其他义项类型的拟合结果良好,其不影响整体的分布情况,多义词模体中多义词的分布符合负二项分布的假设得到证实。

至此,我们考察了多义词高阶长度模体的基本计量特征,即其符合混合负二项分布。这一宏观特征在英汉语中是共同的,但从微观的角度审视,可以发现两种语言中与多义词长度模体相关的差异。首先,就高频多义词模体而言(见表 4 和表 5),英语中长度为 1 的模体占多数,而汉语中长度为 2 的模体在数量上有优势。其次,汉语多义词模体的平均长度(APL = 2.558 5)大于英语(APL = 1.97)。就低频多义词模体而言,以频次小于等于 10 的为例来说,两种语言的情况也基本类似,但数值差异性更显著(汉语的 APL = 13.623 1,英语的 APL = 7.264)。这些低频的多义词模体在数量上约占汉语总模体数量的 10%,在英语中则占到了近 34%。最后,两种语言中的最长多义词模体也有所不同。英语中最长的模体长度为 38,其频次为 39,汉语中最长的是 73,仅出现 1 次。这些差异,有的将在第四节作语言学上的解释,其余的则有待进一步研究。

四　基于多义词模体的英汉语言特征对比分析

作为一项计量和对比研究,本文有两个主要的研究目标:一是探索人类语言中多义词模体的基本统计特征,二是通过多义词模体揭示语言中与多义词相关的本质属性。多义词模体提供了一种可行的方法和一个有效的手段来观察、认识和理解语言,但是,基于多义词模体究竟可以获知哪些相关的语言本质属性这一问题,我们尚未述及。以下是我们为回应此问题所做的一些努力。

基于前述对英汉语多义词的计量考察,我们认为,英汉两种语言中的词在组合关系上具有显著的灵活度和上下文依赖度的区别。整体而言,与汉语多义词相比,英语多义词的意义具有相对较高的灵活度。关于这一点,可以找到三个有力的证据。最重要的证据来自平均义项值。英语多义词的平均义项值在静态的词典和动态的话语中都大于汉语的多义词。对两种语言在词汇语义演变方面的历史考察也可以判断得知,在意义上,英语的词相对开放,而汉语的词则相对保守。其次,英语多义词模体与汉语相比类型更多,具有更大

的变化性,这是多义词多样化的一个表现。另外,英语的多义词模体在数量上多于汉语,在平均长度上小于汉语。我们认为,词义的灵活性与语言的另一属性——复杂性(complexity)相关。

另一方面,英语多义词比汉语多义词具有更大的上下文依赖性。我们深入考察了两种语言中使用频次居前 20 位的多义词模体,获得了数据上的支持。汉语中这些高频多义词模体的长度绝大多数为 2 或 3,几乎全部起始于"1",表明多义词总是跟随在单义词后;类似的,绝大多数的多义词后紧跟着一个单义词,形成"单义词-多义词-单义词"的序列结构。然而,英语中的情形却不相同:绝大多数的高频多义词模体长度为 1,多义词后常为单义词,其前则一般是一个义项值更大的多义词,形成"多义词-多义词-单义词"的组合。在语境中,词的意义是确定、具体、单一的,不管一个词有多少个义项,在一定的语境中通常只使用一个义项。换句话说,语言使用者心理上对词义静态、潜在的多项选择问题通过词与词之间动态、现实的组合得到了解决。对语言使用者而言,尽管在语境中消除词的多义性是一个极其迅速的心理过程,在理解语句的确切意义时,多义词与多义词的组合所造成的理解障碍仍然是一个需要耗费更多时间去解决的问题。在这种情况下,语境是语言使用者寻求帮助的最主要的信息来源,通过不断扩展上下文的窗口范围,才能获得对当前语句的准确理解。从这个意义上说,我们认为,英语多义词具有比汉语多义词更大的上下文依赖度。这一发现有望为二语习得、语言教学、歧义消解、词典编撰等领域的研究带来有益的帮助。

结语

语言模体是计量语言学领域提出的一个新概念,是从组合关系角度对语言的序列结构进行计量考察的一个新的"语言单位"。本文以多义词为研究对象,对英汉语中的多义词模体进行了一些基本的数据统计,探索了两种语言中多义词模体的计量特征,在此基础上,讨论了英汉语中与多义词相关的语言个性特征。

本文的研究表明:(1)英汉语多义词模体的 rank-frequency 分布均符合齐普夫—曼德尔布罗特分布,但分布参数各有不同;(2)英汉语多义词模体的长度模体分布均遵从混合负二项分布;(3)对多义词模体的长度模体的微观考察揭示了英汉语的一些个性特征;(4)英语多义词在意义上比汉语多义词具有更高的灵活度;(5)英语多义词在理解上比汉语多义词具有更大的上下文依赖度。

本文在语言模体的框架下,首次研究了语言的多义词模体。毫无疑问,本文留下的未解问题和尚未涉及的研究议题很多,篇幅所限,语言模体及其相应研究方法不能展现全部效能。未来的研究仍将围绕多义词模体与传统语言学中的一些概念,如词长、词性、句法位置等的相互关系来展开。

Quantitative Properties of Polysemy Motifs in Chinese and English

YANG Jiang[1] & ZHAO Han-bing[2]

(1. School of Foreign Studies, Hunan University of Science and Technology, Xiangtan 411201, China;
2. School of Humanities, Hunan University of Science and Technology, Xiangtan 411201, China)

Abstract: Polysemy is a fundamental and universal property of human languages. This paper explores the basic statistics and fundamental quantitative properties of the polysemy motifs in English and Chinese, on the basis of which the language-specific properties relating to polysemy are discussed. Results show that the English and Chinese rank-frequency polysemy motif distributions fit the Zipf-Mandelbrot distribution with different parameters; the length motif distributions of the English and Chinese polysemy motifs conform to the mixed negative binomial distribution; the micro analysis of the lengths of the English and Chinese polysemy motifs results in revealing many language-specific characteristics; the meanings of English polysemes are of relatively high flexibility compared with Chinese language, and the English polysemes are more context-dependant than the Chinese ones.

Key words: polysemy motif; quantitative properties; distribution; length; Chinese and English

(责任校对　游星雅)