法语料库文本语义接受度评价研究

杜家利[1],于屏方[2]
DU Jia-li[1],YU Ping-fang[2]

1.鲁东大学 外国语学院,山东 烟台 264025
2.鲁东大学 汉语言文学院,山东 烟台 264025
1.School of Foreign Languages,Ludong University,Yantai,Shandong 264025,China
2.School of Chinese Language and Literature,Ludong University,Yantai,Shandong 264025,China
E-mail:dujiali68@yahoo.cn

DU Jia-li,YU Ping-fang.Research on evaluation of semantic accessibility scale based on French corpus.Computer Engineering and Applications,2009,45(29):125-127.

Abstract: The study on Semantic Accessibility Scale(SAS) based on French corpus comes under the umbrella of corpus-involved SAS research.Systematic random sampling skill is used here to discuss the comparison of six different groups extracted from one text.The formula,$A+BX\leqslant C$,is provided to equidistantly extract L'Assommoir(1953 ed.).“A”means the start page;“B”,the space of extraction;“C”,the total number of text;“X”,the set of available pages.If the condition is $B\in(10;5;4;3;2;1)$,sampling ratios(SR)are 9.55%,20.29%,25.13%,33.13%,50.36%,100%,correspondingly bringing different values of SAS,namely,0.089 7,0.084 1,0.085 4,0.084 7,0.084 8,0.085 4.The non-relevance between SR and SAS establishes the fact that the system of SAS evaluation based on English and Japanese corpora can safely extend to French corpus,which helps the critics to analyze French texts by computer.
Key words: literary texts;French corpus;semantic accessibility scale;sampling ratio

摘 要:法文本语义接受度(SAS)研究属于基于语料库的SAS研究分支。以等距离系统随机抽样方法进行对比实验。抽取公式为$A+BX\leqslant C$,A为起始页码,B为抽取间距,C为文本总页码,X为可取页数集。当以L'Assommoir(1953版)为语料,$B\in(10;5;4;3;2;1)$时,词句抽取率(SR)为9.55%,20.29%,25.13%,33.13%,50.36%和100%,SAS为0.089 7,0.084 1,0.085 4,0.084 7,0.084 8和0.085 4。依次攀升SR没有带来SAS显著变化。说明基于英日语料的SAS评价体系也适用于法文本,便于文学评论家对法文本进行量化评析。
关键词:文学文本;法语语料库;语义接受度;抽取率
DOI:10.3778/j.issn.1002-8331.2009.29.037 **文章编号**:1002-8331(2009)29-0125-03 **文献标识码**:A **中图分类号**:TP311

1 引言

语义接受度(Semantic Accessibility Scale,SAS)研究是针对语料库文本进行的篇章难易度分析,单语和跨语种语料库基础研究为法语料库文本SAS评价提供了理论条件。

现行单语语料库研究多侧重设定语素、词、句封闭域的语音、语法、语义研究。如统计分析在单语语料库中的实用[1];自动文摘系统研究[2];面向科技摘要的概念判定和辨析[3];手势、唇形和嗓音合成的语音器[4]和文本视听合成[5]研究;基于语料库的质控网关的自动标引[6]、话语分析[7]、语音变异[8]和话对语义偏好[9]研究;以及语料库建构下写作技能[10]、独立词变量缺失[11]研究等。

跨语言语料库多关注多语言的系统对比和对照。如英日汉语料的语用对比[12];日意幼儿习得词序对照[13];多语言增量聚类的解歧策略[14];日汉平行语料库自组织语义图谱词对齐研究[15];基于英法对比语料库的跨词阅读[16]、在线异步讨论[17]、学术期刊摘要[18]、话语再现的语用功能分析[19]研究;英汉[20]、英日汉[21]的标题对齐和对译研究;基于英法挪威语言的学术文本中语言文化关联性研究[22]等。

随着网络发展,关注篇章层面的文本语料库研究正成为语言学和文学评论的交叉热点。如文本风格研究[23]、文本难易度研究[24]等。

该文属基于单语料库的文本可理解程度(SAS)研究,主要探讨法语料库中抽样文本和理解程度的关联性。此研究主要分两类:横向对比不同作者创作的文本间[25]和纵向对比同一作者文本不同抽取率间[26]的SAS研究。该文侧重后者,并验证英日语料库的SAS评价公式在法语中的适用性。

2 现行语义接受度评价研究—基于英日语料库的分析

英日语料库SAS评价研究来源于对现行机器语义评价体

基金项目:国家社会科学项目(No.08BYY046);山东省社会科学规划项目(No.07CWXJ03)。
作者简介:杜家利(1971-),男,讲师,研究方向:计算语言学;于屏方(1971-),女,博士,讲师,研究方向:应用语言学。
收稿日期:2009-05-18 **修回日期**:2009-06-25

系和自动文摘系统的借鉴,是多参数融合的变量分析法,是在召回率和精确度分析法、F-Measure 测试法、Rouge 和 F-New-Measure 分析法基础上提出的文本可理解程度的分析策略。测试结果强调自动性、复现性和模式性,即评价结果的自动生成,评价量值的复现回归,评价程序的模式化运行。

文本评价的复杂性决定英日 SAS 评价系统的多参数性,涉及取样文本的词句长和抽取率。设单位句子含词量为句长 L、百词中超常用词量为词长 H、词句长之和的加权值为 0.4,文本取样句数 $S1$、取样词数 $W1$、文本总句数 S,总词数 W,句抽样率为 $P1$,词抽样率为 $P2$,词句综合抽取率 SR,语义接受度为 SAS。具体公式如下:

$$P1=\frac{S1}{S} \tag{1}$$

$$P2=\frac{W1}{W} \tag{2}$$

$$SR=\frac{2\times P1\times P2}{P1+P2}=\frac{2\times S1\times W1}{S1W+SW1} \tag{3}$$

$$SAS=\frac{1}{0.4(L+H)}\times\frac{P2}{P1} \tag{4}$$

$$SAS=\frac{1}{0.4(L+H)}\times\frac{S\times W1}{S1+W} \tag{5}$$

英语属屈折语系,H 指百单词中三音节及以上单词的数量。日语属粘着语系,H 指百词中所含五音拍及以上词总数。当英语料库取样文本 SR 为 9.86%,20.17%,24.99%,33.62%,49.90%和 100%时,各组对应 SAS 为 0.151 8,0.147 9,0.147 2,0.144 8,0.148 8 和 0.148 4。当日语料库取样文本 SR 为 9.11%,19.29%,25.35%,32.89%,50.97%和 100%时,各语义接受度为 0.085 2,0.089 3,0.089 1,0.089 3,0.088 4 和 0.088 1。基于英日语料库的文本 SAS 研究得出两个结果:(1)SR 依次攀升未导致 SAS 值显著波动,证明 SAS 公式可用于英日文本的语义理解度评价;(2)抽取率大致相当时,英日抽样文本语义理解度偏差较大,说明两抽样文本具有不同的写作风格[25]。

3 屈折语特点及法文本 SAS 评价标准设定

屈折语是比较语言学根据结构标准建立的语言类型之一,词用形态表示语法关系:一般包含不止一个语素,但不像粘着语(如日语)那样语素和语子的线性序列之间有一一对应关系,也不像孤立语(如汉语)那样都是不变形词而且句法关系主要靠词序表示。大体上说,英语和法语都属于屈折语范畴。

与英语类似,法语通常使用的词是三音节以内的词,超出三音节的用词量可作为文本理解难易程度的依据。已通过英日语料库证明的 SAS 公式所涉及的变量意义在下文中相同,其中,H 为百单词中三音节及以上单词的数量。

法语料库文本取样以荣膺国葬、在法国文学家享誉盛名、现实主义作家 Emile Zola(左拉)之作 L'Assommoir《小酒店》为语料(Larousse 公司 1953 出版),该文以法国资本主义萌芽和手工业没落的 1860 巴黎为背景,描写了一对工人夫妻如何在工伤后懒散沦落、因酗酒而贫困潦倒直至发疯死亡的故事。

下文主要讨论基于此语料的多抽取率和相应语义值的关系,验证适用于英日语料库的 SAS 评价系统是否在法语料库中具有可行性。

4 基于法语料库的文本语义接受度分析

法语料库文本 SAS 研究采用等距离系统随机抽样方法进行对比实验。抽取公式为 $A+BX\leqslant C$,A 为起始页码,B 为抽取间距,C 为文本总页码,X 为可取页数集。其中起始页码的选定按照随机量表选择。例如 $A=7$,$C=500$,$B=10$,则 $X\in(0,49)$,可解释为随机初始页码为第 7 页,抽取间距是 10 页,而且总长度是 500 页的取样文本,共可以抽取总数 50 页的取样量进行研究。讨论当 $B\in(10;5;4;3;2;1)$时,依次求得 SR 与对应 SAS 的值。

4.1 法语料库抽样文本多抽取率对照

L'Assommoir 文本中 $C=487$,取样分六组进行对照:Group 1:$A=6$,$B=10$,$X\in(0,48)$;Group 2:$A=4$,$B=5$,$X\in(0,96)$;Group 3:$A=3$,$B=4$,$X\in(0,121)$;Group 4:$A=2$,$B=3$,$X\in(0,161)$;Group 5:$A=1$,$B=2$,$X\in(0,243)$;Group 6:$A=1$,$B=1$,$X\in(0,486)$。

第一组抽取 49 页:6,16,26,36…466…486。第二组抽取 97 页:4,9,14,19…469…484。第三组抽取 122 页:3,7,11,15…471…487。第四组抽取 162 页:2,5,8,11…470…485。第五组抽取 244 页:1,3,5,7…471…487。第六组全文统计 487 页:1,2,3…487。

如表 1 和图 1 所示,法文本抽取间距 $B\in(10;5;4;3;2;1)$时,抽取率 SR 依次为 9.55%,20.29%,25.13%,33.13%,50.36%和 100%,抽取率为依次攀升的曲线,抽取间距越大,抽取率越小,抽取间距引发抽取率规则性变化。

表 1 L'Assommoir 对照组抽取率数据列表

类别	P49	P97	P122	P162	P244	P487
$W1$	16 887	34 589	43 122	56 912	85 984	170 918
$S1$	754	1 658	2 041	2 688	4 110	8 154
W	170 918	170 918	170 918	170 918	170 918	170 918
S	8 154	8 154	8 154	8 154	8 154	8 154
SR	0.095 5	0.202 9	0.251 3	0.331 3	0.503 6	1.000 0

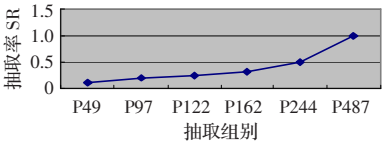

图 1 L'Assommoir 对照组抽取率变化图

4.2 法语料库抽样文本多语义接受度对照

抽取率获值后,可计算出单位句子含词量(句长 L)和百词中超常用词量(词长 H),再利用公式求得各 SAS。

由表 2 和图 2 所示,各 SAS 值为 0.089 7,0.084 1,0.085 4,0.084 7,0.084 8 和 0.085 4。$SAS\in(0.084\ 1,0.089\ 7)$。当 $B=10$ 时,SAS 值偏离均值较大;$B\in(5;4;3;2;1)$时,SAS 呈现不规则变化,但围绕均值波动。

表 2 L'Assommoir 对照组语义接受度数据列表

类别	P49	P97	P122	P162	P244	P487
L	22.92	22.29	22.23	22.39	22.09	21.99
H	6.87	7.31	7.29	7.42	7.35	7.27
SAS	0.089 7	0.084 1	0.085 4	0.084 7	0.084 8	0.085 4

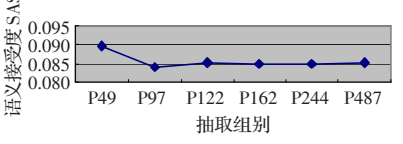

图 2 L'Assommoir 对照组语义接受度变化图

4.3　法语料库抽样文本语义接受度与抽取率对照

如果抽样率规则变化能带来语义接受度的规则变化，则说明 SR 应该是 SAS 评价体系的参量，不同比率抽样会带来语义评价值的不同，文学评论者就不能采用随机抽样的方式进行文本语义分析。相反，如能证明两者不具有关联性，分析者就可采用抽样调查的方式进行文本语义的风格研究。

由表 3 和图 3 所示，抽取率依次为 9.55%，20.29%，25.13%，33.13%，50.36%和 100%时，语义接受度为 0.089 7，0.084 1，0.085 4，0.084 7，0.084 8 和 0.085 4。语义值基本保持平衡，依次攀升 *SR* 没有带来 *SAS* 显著变化。*SAS* 围绕均值波动，不随 *SR* 递增而变化。此曲线说明多样抽取率不会或基本不会带来特定文本语义值的偏差，适用于英日语料库文本的语义接受度评价公式可扩展到法语料库文本，便于文学分析者根据抽样的单位词句长和超常使用的词数进行文本语义分析。

表 3　L'Assommoir 对照组抽取率与语义接受度数据列表

类别	P49	P97	P122	P162	P244	P487
SR	0.095 5	0.202 9	0.251 3	0.331 3	0.503 6	1.000 0
SAS	0.089 7	0.084 1	0.085 4	0.084 7	0.084 8	0.085 4

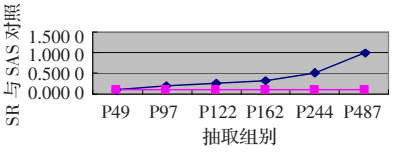

图 3　L'Assommoir 抽取率与语义接受度对比图

5　结论

基于语料库的文本语义接受度(SAS)研究是计算语言学与文体学的交叉。以法语料库文本 L'Assommoir 为语料，利用英日语料库验证后的 SAS 评价体系对其进行文学文本可理解程度的量化分析。

实验证明：(1)抽样间距与抽取率规则变化，间距越大，抽样率越小；(2)抽样率规则变化未引起代表文本风格特点的语义接受度的规则变化；(3)抽样率不是 SAS 公式的变化参数，但大间距抽样会导致分析值偏离 SAS 均值较大；(4)涉及抽样文本词句数、单位词句长、超常使用三音节及以上词数的 SAS 公式对法语文本具有敏感性，可用于分析法语料库文本的语义接受度。

尽管适用于英日语料库的 SAS 独立于 SR 的实验在法语文本中也得到验证，但存在于英日实验中的较大抽样间距(B=10)带来 *SAS* 值较大偏离的现象在文中也再次出现，此类循环出现的数据失真是否可控以及是否具有"型"的特征值得进一步探讨。

致谢：鲁东大学外国语学院法语系刘媛媛老师的数据支持。

参考文献：

[1] Sakamoto K，Terai A，Nakagawa M.Computational models of inductive reasoning using a statistical analysis of a Japanese corpus[J].Cognitive Systems Research，2007，8：282-299.

[2] 傅间莲，陈群秀.一种新的自动文摘系统评价方法[J].计算机工程与应用，2006，42(18)：176-177.

[3] Saggion H，Lapalme G.Concept identification and presentation in the context of technical text summarization[C]//Proc of the Workshop on Automatic Summarization.New Brunswick，New Jersey：Association for Computing Linguistics，2000：1-10.

[4] Attina V，Beautemps D，Cathiard M A，et al.A pilot study of temporal organization in cued speech production of French syllables：Rules for a cued speech synthesizer[J].Speech Communication，2004，44：197-214.

[5] Benoît C，Goff B L.Audio-visual speech synthesis from French text：Eight years of models，designs and evaluation at the ICP[J].Speech Communication，1998，26：117-129.

[6] Névéol A，Rogozan A，Darmoni S.Automatic indexing of online health resources for a French quality controlled gateway[J].Information Processing and Management，2006，42：695-709.

[7] Morlec Y，Bailly G，Aubergé V.Generating prosodic attitudes in French：Data，model and evaluation[J].Speech Communication，2001，33：357-371.

[8] Adda-Decker M，de Mareüil P B，Adda G，et al.Investigating syllabic structures and their variation in spontaneous French[J].Speech Communication，2005，46：119-139.

[9] Léon J.Preference and "bias" in the format of French news interviews：The semantic analysis of question-answer pairs in conversation[J].Journal of Pragmatics，2004，36：1885-1920.

[10] O'Sullivan Í，Chambers A.Learners' writing skills in French：Corpus consultation and learner evaluation[J].Journal of Second Language Writing，2006，15：49-68.

[11] Armstrong N.Variable deletion of French ne：A cross-stylistic perspective[J].Language Sciences，2002，24：153-173.

[12] Clancy P M，Thompson S A，Suzuki R，et al.The conversational use of reactive tokens in English，Japanese，and Mandarin[J].Journal of Pragmatics，1996，26：355-387.

[13] Gervain J，Nespor M，Mazuka R，et al.Bootstrapping word order in prelexical infants：A Japanese-Italian cross-linguistic study[J].Cognitive Psychology，2008，57：56-74.

[14] Lee K S，Kageura K，Choi K S.Implicit ambiguity resolution using incremental clustering in cross-language information retrieval[J].Information Processing and Management，2004，40：145-159.

[15] Ma Q，Kanzaki K，Zhang Y，et al.Self-organizing semantic maps and its application to word alignment in Japanese-Chinese parallel corpora[J].Neural Networks，2004，17：1241-1253.

[16] Pynte J，Kennedy A.An influence over eye movements in reading exerted from beyond the level of the word：Evidence from reading English and French[J].Vision Research，2006，46：3786-3801.

[17] Lewis D M.Arguing in English and French asynchronous online discussion[J].Journal of Pragmatics，2005，37：1801-1818.

[18] Bonn S V，Swales J M.English and French journal abstracts in the language sciences：Three exploratory studies[J].Journal of English for Academic Purposes，2007，6：93-108.

[19] Perrin L，Deshaies D，Paradis C.Pragmatic functions of local diaphonic repetitions in conversation[J].Journal of Pragmatics，2003，35：1843-1860.

[20] Yang C C，Li K W.Building parallel corpora by automatic title alignment using length-based and text-based approaches[J].Information Processing and Management，2004，40：939-955.

[21] Shimizu T，Ashikari Y，Sumita E，et al.NICT/ATR Chinese-Japanese-English speech-to-speech translation system[J].Tsinghua Science and Technology，2008，13(4)：540-544.

(下转 179 页)