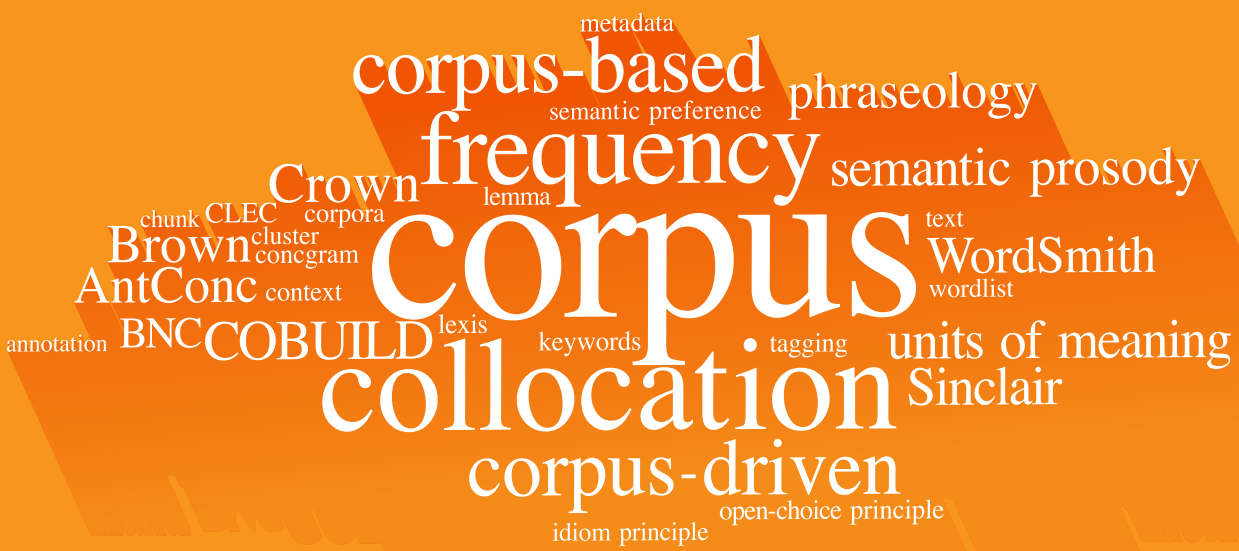
《中国学术期刊网络出版总库》及CNKI系列数据库入选期刊

语料库语言学

CORPUS LINGUISTICS

2
Vol. 5 No. 2
第5卷 第2期
2018

北京外国语大学中国外语与教育研究中心
中国英汉语比较研究会语料库语言学专业委员会
许家金 主编



外语教学与研究出版社
FOREIGN LANGUAGE TEACHING AND RESEARCH PRESS

语料库语言学
（半年刊）

Corpus Linguistics
(Biannual)

主　管：中华人民共和国教育部
主　办：北京外国语大学
承　办：中国外语与教育研究中心
　　中国英汉语比较研究会
　　语料库语言学专业委员会
出　版：外语教学与研究出版社

Administered by the Ministry of Education of China
Directed by Beijing Foreign Studies University
Edited at the National Research Centre for Foreign Language Education and Corpus Linguistics Society of China, China Association for Comparative Studies of English and Chinese
Published by Foreign Language Teaching and Research Press

主　编：许家金
编　校：华　雨　徐秀玲

Editor: Xu Jiajin
Proofreaders: Hua Yu and Xu Xiuling

电　　话：（010）88816828
电子邮箱：bfsucrg@sina.com
投稿网址：http://ylyy.chinajournal.net.cn

本刊地址：北京市西三环北路19号北京外国语大学中国外语与教育研究中心
《语料库语言学》编辑部（100089）

《语料库语言学》
2018年　第5卷　第2期

目　录

学者聚焦

研究论文

研制开发

书刊评介

CORPUS LINGUISTICS

Volume 5, Number 2, 2018

Table of Contents

An interview with Hilde Hasselgård

University of Oslo, Norway
13:30-15:15, 27 June, 2018

(**XLX:** Xiuling Xu; **HH:** Hilde Hasselgård)

XLX: When did you initially know about corpora and corpus linguistics?

HH: I had Stig Johansson as my teacher already when I was a master student in the late 1980s. So I knew about corpora at that stage and learned how to search in a corpus. But I didn't use corpora for my MA thesis although I did use methods from corpus linguistics only because it was manual. I didn't own a computer when I was an MA student, and we had only two computers that MA students could use and of course no Internet so it was quite difficult to use corpora in practice. But when I started my PhD, I started using corpora.

XLX: Who has the greatest impact on you regarding corpus linguistics? And in what ways?

HH: That is definitely Stig Johansson, who was my teacher and supervisor both for my MA and PhD, and later he was a colleague and a friend. He was one of the pioneers in English corpus linguistics overall, so having him close by was very important. He took me to my first ICAME conference in my first year as a PhD student in 1990 when the ICAME conference was still very small. So he also introduced me to corpus linguists.

XLX: What do you think are the most fascinating aspects of corpus linguistics?

HH: I think it's great to have authentic data that we can access very easily.

For people who study a language that isn't their first language, it's a great advantage to investigate authentic language instead of relying on intuition because people generally have poor intuitions. Sometimes you find really surprising things that you wouldn't have thought of yourself. When you've carried out a corpus linguistics investigation, it should be replicable if you have specified how you searched and how you analysed and sorted the material. So it is good scientific practice as well. We can also use corpora in teaching. When corpus linguistics methods are used in a simple manner, students can do the same kind of things as we do as researchers. I think that's an excellent thing about corpus linguistics. You can do it at several levels of expertise.

XLX: How do you see the synthesis of linguistic theory and corpus linguistics?

HH: I do not think that corpus linguistics is a theory and I suppose most people do not think that. In corpus linguistics, you need a linguistic theory in order to interpret your findings and perhaps even to define your research questions. I think several theories go very well with corpus linguistics, but they would tend to be functionally oriented theories that aim to account for language use rather than theories that aim to model how language is produced. I think that you do need linguistic theory in order to make sense of corpus linguistics and particularly now that the technical bits are very easy compared to when I got to corpus linguistics. Often people spent much time and energy on the technical aspects and were not so interested in theory. They were so fascinated with the fact that they could look through a million words in a few hours instead of collecting examples for years.

XLX: Are there contrastive linguistic issues that can be significantly better addressed with the assistance of corpus linguistics?

HH: What contrastive linguistics gained from the multilingual corpora was a much better empirical basis for carrying out what both Stig Johansson and Carl James refer to as systematic comparison of two languages. Previously I suppose people had compared grammars, and to some extent vocabularies, but it was much more difficult to know what to compare, particularly beyond

rather trivial topics such as how you form a question in English compared with Norwegian, Spanish or Chinese. You can do much more sophisticated things with a corpus. You can study translation correspondences, which translation scholars obviously had done, but they had been looking at one text at a time because that's what you can do without a corpus.

Another thing is that parallel corpora help to define a *tertium comparationis* (TC) for contrastive analysis. TC is about a background of sameness against which two linguistic items can be compared, and this is something that is built into a bidirectional corpus model such as the English-Norwegian Parallel Corpus. This corpus has originals in both English and Norwegian and translations into the other language. We know that translators try their best to preserve the meaning and function of the original, so we can regard the translation relation as a TC. You can check translation correspondences in both directions and be fairly certain that you are comparing like with like. That is a huge advantage.

XLX: Right. Before the use of corpora, it was very difficult to define TC precisely. People had to rely on intuition.

HH: Obviously people could rely on dictionary equivalents if they were comparing vocabularies. What else? Grammatical functions I suppose. For example, you could decide to study modality by comparing the modal auxiliaries in two languages. But with a bidirectional translation corpus, you can search for a modal auxiliary, for example *should,* to see how it is reflected in the other language. The translations may be not just modal auxiliaries but also adverbs, adjectives and nouns, so you get a much broader picture of modal meaning and how it is expressed across languages.

XLX: To what extent can corpus linguistics enhance our understanding of some theoretical linguistic issues?

HH: I think the example of modality is a very good illustration, because it is a semantic field that contains much more than the modal auxiliaries. I'm sure there are other fields like that and where languages can have or prefer

different resources to express the same kind of meaning. You get that even with languages that are typologically rather similar, like the Germanic languages. But I imagine that if you want to look at time reference for example, and you compare English with Chinese, which doesn't have tense, you will find a spread of devices in Chinese that you can use instead of verb morphology to express present time reference. Sometimes there will be no marker because people just know that you are talking about the present time. The main thing is that the English tense system would correspond to something else in Chinese, and the parallel corpus can help you see what that is.

XLX: Do you think corpus linguistics has come of age as a discipline of its own?

HH: That's an interesting question, because in one way it has and in another way I think it's still developing. Maybe people felt it was more of a field when the discipline was young. Then people started to think that they study linguistics by means of corpora. So they came to regard corpus linguistics more as a method for doing linguistics. From what I've seen at recent ICAME conferences, where you can kind of monitor the developments, I think it's still a field but it's getting diversified. We now have subfields, such as corpus pragmatics or learner corpus research. The field diversifies and it also kind of sneaks into other disciplines. I mean there used to be a water-tight division between what we call theoretical linguistics and corpus linguistics, but now theoretical linguists use corpora to get material and corpus linguists use theory to interpret their material.

XLX: Can you comment on the corpus research at the University of Oslo from an insider's perspective? e.g. What are the main strands and characteristics of corpus research at the University of Oslo? What are the key contributions of corpus research at the University of Oslo to linguistics research in general?

HH: Our university has a Text Laboratory (in the Department of Linguistics and Scandinavian Studies) where they have developed corpora mainly of Norwegian but also other languages, and I'll come back to their work. It's

in the English Department that corpus linguistics has the longest history at the University of Oslo. That was again because of Stig Johansson who introduced corpora already at the end of 1970s. I think at that time only he used them. Stig's pioneering role started when he spent a sabbatical in Lancaster and met Geoffrey Leech, who was struggling to compile the corpus that was to become LOB (LOB is short for Lancaster, Oslo and Bergen). Stig offered to help with the project that seemed to be stopping before it was finished. Knut Hofland in Bergen agreed to do the technical stuff, and together they managed to complete the LOB Corpus. As you know, the LOB Corpus was only the second computerized corpus of English after Brown. These were the corpora that people used for rather a long time. Obviously they were extremely important in establishing corpus methodology. The second major contribution was also Stig Johansson's work: it was the multilingual development with the English-Norwegian Parallel Corpus. The idea was first presented at an ICAME conference in 1993. Already in the mid-1990s we had materials to work with and we had our first symposium in 1994 in Lund with colleagues from Sweden, Denmark and Finland. But the ENPC was the first parallel corpus that was made public and available to a wider circle of people than those who were directly involved in the project. It contributed a lot to parallel corpus methodology too. There have been many other corpus projects that have been based on the model of the ENPC. The first were the sister projects in Sweden and Finland. The Finnish corpus was I think only partly completed, and the Danes decided not to join, unfortunately. There are also similar corpus projects elsewhere with other language pairs, which relate directly to what was started here in the 1990s.

XLX: How do you view your own role in corpus research at the University of Oslo and your contribution to corpus linguistics beyond UiO and Norway?

HH: Within the English Department, I suppose that both Signe Ebeling and myself see ourselves as Stig's children in our teaching and research. We try to continue the tradition. Obviously the field has moved forward, so we can't do the same things as Stig did. We have to do something new. But we had a good starting point because Stig introduced us to his network of

corpus linguists. I think when we work with English in Norway or another country where English is a widely taught and very dominant second language, contrastive linguistics is a natural field to work within. Another is English as a foreign or second language. And this is something that we can do well in a non-English speaking country. One international role is that I have been on the board of ICAME and am now vice president in the Learner Corpus Association. Also, when you know people in the field, you are asked to be on committees and to review projects, articles, and books for example. So I guess that is another role: participating in the international community of corpus linguistics. I have published internationally, so there are people outside Norway who have read about my research. I have also edited a number of publications, including *Languages in Contrast*, which is an important journal for contrastive linguistics, and we get many corpus-based submissions.

XLX: You've talked about the history of corpus research at the University of Oslo. How do you see the future directions of corpus research at the University of Oslo?

HH: I hope it will continue. Of course we give courses on corpus linguistics as an obligatory part of our MA program in English linguistics because we believe it's a useful thing for people to know. Hopefully there will be staff here after my time as well who can continue the tradition. We also have colleagues who work with historical corpora and others who work with languages other than English. Our future directions will depend on what personnel we have and their research interests. But it would be great to continue with more contrastive research, with corpora that are better in the sense of newer and more comprehensive. There are very natural limitations because not everything is translated. Particularly from Norwegian into English you don't get a lot of published translation. We have been looking at ways of using comparable corpora in a more productive way. The main problem is the *tertium comparationis* I was talking about earlier, but maybe if we use comparable and translation corpora in combination with each other, then we might find a reliable method of comparing data from comparable corpora as well. I think that's an exciting avenue to pursue. Another thing that we could

study is the way that English is spreading in Norwegian society, outside of the educational institutions, more or less as a second language rather than a foreign language. I think it would be very interesting to look at English in workplaces for example. It will be different from both learner English and English as a lingua franca in the sense that the users of English won't be in a learning situation, and it will be within predominantly Norwegian contexts. It would be possible and very interesting to try to capture some of this development in corpora.

I mentioned in connection with a previous question that there is a Text Laboratory at the University of Oslo, where people work mainly with Norwegian and other Scandinavian languages. An interesting thing that they've done is to collect spoken data from various kinds of linguistic communities. So they have developed corpora of spoken language that are very well suited for studying variation, such as dialect corpora. They have also compiled corpora for lexicographical purposes. Hopefully they will continue to develop ways of preparing and displaying spoken material for corpus use.

XLX: What are the main strands of Scandinavian Corpus Linguistics?

HH: I'll focus on English corpus linguistics in the Scandinavian countries Norway, Sweden, Finland and Denmark. Sweden has been very important, with their very early pioneer Jan Svartvik, who was one of the authors of the Quirk *et al*. grammars. He got interested in corpus linguistics at University College London where he and Quirk developed the London Lund Corpus, which was a pioneering step in spoken corpus research. The collection and much of the annotation took place in London but the computerization I think was mainly in Lund. The Swedes were very early to do research on spoken English and they had a research project about differences between speech and writing in the 1980s. The Finns have done tremendous work in historical corpus linguistics. Matti Rissanen was behind the first computerized historical corpus of English, the Helsinki Corpus of English Texts. That was pioneering work. His team had to find ways of representing Old and Middle English in computer readable format and also ways of standardizing texts so they would be searchable. Later, the Helsinki group has developed a lot

of different historical corpora, for instance specialized corpora of letters, medical texts and so on. Unfortunately, Denmark doesn't seem to have a strong tradition of English corpus linguistics, but there used to be people who were interested in corpora of English for specific purposes. To simplify, maybe the contributions from Scandinavia are specialized branches of corpus linguistics, so spoken English in Lund, multilingual corpora here and historical corpora in Finland.

XLX: Could you refer me to some of the pioneering and leading Scandinavian corpus linguists in these branches of linguistics?

HH: In terms of the spoken language, I already mentioned Jan Svartvik. There are also some other people from Sweden, for example Bengt Altenberg, Karin Aijmer, Gunnel Tottie and Anna-Brita Stenström. All these people have now retired, but Karin Aijmer is still very active. She has also done work in the new field of corpus pragmatics, and I would say she has contributed to developing new methods and using corpora in new ways. The Norwegian corpus linguist Gisle Andersen has also done important work in corpus pragmatics.

XLX: How about the leading corpus linguists in contrastive linguistics?

HH: We can't get around Stig Johansson. Jarle Ebeling and Signe Ebeling worked with him on technical and practical matters. They've been much more hands-on with the corpus development than me. I wasn't directly involved in the work, although I was in the department at that time and was part of the ENPC research group in a way. But of course I have had the advantage of using the corpus in my research. I should also mention Cathrine Fabricius-Hansen, who has worked with the Oslo Multilingual Corpus, mainly with a focus on German.

In Sweden, Bengt Altenberg and Karin Aijmer have done very important contrastive work, and they have also studied learner language. Jan Svartvik was never part of the parallel corpus project, but has been influential for English Corpus Linguistics overall, and he had a research group in the 1970s

that worked with learner language analysis and contrastive topics. They used a lot of empirical material, but it was not computerized. I think this early empirical work, without corpora, was part of the reason why it was so easy to get people in Lund interested in contrastive corpus linguistics, because they were already interested in similar research questions. And they saw that a parallel corpus was an excellent way of studying language contrasts. A fun-fact about the project that produced both the English-Norwegian and the English-Swedish Parallel Corpus is that Stig Johansson, Bengt Altenberg and Karin Aijmer were old friends from Lund, and that was a very good starting point for successful cooperation.

XLX: Could you name some historical corpus linguists? Is it mainly because of Matti Rissanen that Finland has a strong tradition of corpus-based historical linguistics?

HH: Matti Rissanen is obviously an important name. He developed the Helsinki Corpus and managed to get a large research group together. Many of these people are also important historical linguists, such as Terttu Nevalainen and Irma Taavitsainen. Merja Kytö was Matti's research assistant on the Helsinki Corpus and she later moved to Uppsala in Sweden. We also have historical corpus linguists in Oslo, for example Kristin Bech has had two projects on Old English, and Gjertrud Stenbrenden has done important research in historical phonology based on corpora compiled in Edinburgh.

XLX: Could you share some stories with regard to the interaction of Scandinavian Corpus Linguistics with corpus linguistics in other regions, for example, with Granger's team in Belgium, and how you came into cooperation with them?

HH: Before Karin Aijmer retired, she organized several workshops in Gothenburg on both cross-linguistic and learner corpus research. That was one forum of contact and cooperation. We also cooperate with the University of Louvain. Apart from ICAME conferences, Sylviane Granger took part in the first symposium about parallel corpus research in Lund in 1994. She had started

compiling the ICLE Corpus then and was developing a method for studying learner corpora, which was also in a way contrastive. Now we think of cross-linguistic and learner language research as two different but related fields, but they used to be much more closely connected in people's minds. I have met up with people from Louvain from time to time because we have a lot of shared research interests in both contrastive linguistics and learner corpus research. In Louvain they have a parallel corpus similar to our ENPC, and on their initiative, or invitation, we now have Norwegian components of the learner corpora ICLE and VESPA. Of course once you get interested in learner corpus research and someone from Louvain is interested in cooperation with you, it's a big bonus, because the learner corpus research obviously comes from Louvain-la-Neuve and other people have followed, like the bidirectional parallel corpus comes from here and other people have followed. So I suppose we've always been interested in each other's work and that we can learn from each other.

XLX: What do you consider are some of the essential qualifications for a corpus linguist? e.g. Should a corpus linguist be able to do programming? How important is statistics for a corpus linguist?

HH: I think it's becoming more and more demanding to be a corpus linguist, because the number of things you need to know keeps being added to. I think the most important thing is still the linguistic part. As a corpus linguist who works with a foreign language, you have to be fluent in that language, because otherwise you can't tell what you see, you can't interpret the data. You need to know linguistics, so you know how to categorise things. In addition to those basic skills, you need to know how to handle the corpus and how to construct a search string that will get you the data that you need and not overlook a lot of relevant data. And you need to know corpus properly. For example, you need to know what's in your corpus. There are a surprising number of people who use a corpus but they do not know what they are looking at, whether it's a corpus of fiction, or whether the texts come from the 1930s or the 2000s. As seasoned corpus linguists we know that makes a difference. Diachronic studies have shown that age matters, and results can differ even across short time spans. One of the important

insights from for example Douglas Biber's work, is that genres differ in how they use language. People need to know these things when they use corpora. Because if you don't know what's in the corpus, you can't really use it. As I said, corpus linguistics is diversifying, so what skills you need to have depends on what you want to do. To answer your question "Should a corpus linguist be able to do programming?" Not necessarily, because if you are a corpus user rather than a corpus compiler or software developer, you don't need to do programming.

It's important to understand the quantitative side as well. Of course that part of corpus linguistics has developed immensely in the last few years, and the statistical methods have become more and more sophisticated. I mean it's not just that we now have statistical measures that are more precise than the percentages and normalized frequencies that people used before, but statistics is also becoming a separate specialization. I think it's important for a corpus linguist to have some understanding of statistics and the more you know the better. But I think few people can be a fully skilled statistician and a fully skilled linguist at the same time. It's important to find someone to cooperate with, or to know your limitations so that you don't say things that you don't have evidence for. Linguistics, not programming or statistics, is the main home ground of a corpus linguist. The other things are important but not as essential as the linguistic understanding.

XLX: Do you have any advice to young scholars who wish to do corpus research?

HH: I think for someone who is getting into corpus linguistics now, it is very important to get an overview of the field, and decide on your priorities. Because as the field is diversifying we need more and more skills. I think it's unlikely that you'll get all of those kinds of expertise in one head, so you need to cooperate with people who have the skills that you don't have. So I think I would advise people to find out who they can cooperate with and how.

XLX: Do young corpus linguists need to know the history of corpus linguistics?

HH: Maybe it is not essential, but it's good for them to know it. I think it is useful for people who have grown up with Internet and computers to understand how difficult it was for linguists to get their data without powerful computers, and also to appreciate the kind of work that goes into compiling a proper corpus. If you read an old corpus study that uses the LOB and the Brown corpora, which are one million words with various genres in them, and someone writes about genre variation on the basis of that, maybe you would say "Oh gosh, there is almost nothing in those corpora and they are from 1961. Why should I be interested or how can people say anything on the basis of so little data?" if you don't realize that at the beginning of corpus linguistics a million words was massive. It was really at the limit of what computers could hold and what you could search in, so it was a lot of data. It was much more adventurous to build a one-million-word corpus than it is to build a one-billion-word corpus now. Computer power and storage were very real limitations then and people had to think of clever ways of getting all that data into as little space as possible. I think it's important to understand that people who used small corpora and percentages rather than sophisticated statistics were not stupid. They were people who worked with whatever they had and perhaps even very adventurous pioneers who paved the way for what we can do now.

XLX: I totally agree. Thank you very much.

从态度系统看西方媒体对中国股市形象的构建*

对外经济贸易大学 江进林 张皎皎

提要：本文基于Fairclough的三维分析框架，以评价理论中的态度系统为描写框架，研究西方主流媒体近5年对中国股市的报道，以揭示西方媒体的态度倾向，并分析其态度背后的深层次社会历史原因。研究表明，西方主流媒体大量使用表达消极评价意义的词汇，构建出十分负面的中国股市形象。西方媒体在肯定中国股市取得一定成绩的同时，大力批判中国经济的道德性。此外，西方媒体认为中国股民素质不高，斥责中国股市的管理方式和政府的行政干预手段。这种话语模式反映了西方媒体忽视中国股市的发展阶段、倡导西方“自由市场”体制的用心。

关键词：西方媒体、中国股市、批评话语分析、态度系统、语料库

1. 引言

中国经济的平稳快速发展对世界经济具有巨大的推动力。股市作为中国金融市场的重要组成部分，对中国经济的发展起着重要作用。《中华人民共和国国民经济和社会发展第十三个五年规划纲要》提出，要完善金融机构和市场体系，有效防范和化解金融危机。自2008年国际金融危机爆发之后，西方媒体开始争相报道中国股市的行情。然而，西方媒体视野中的中国股市是何形象？它们是如何塑造这种形象的？为何会这样塑造？学界迄今尚未对此展开实证研究。本文通过自建西方媒体中国股市报道专用语料库，以评价理论中的态度系统为描写层，结合索引行、搭配等语料库研究方法，从批评话语分析的视角研究西方媒体对中国股市的报道，以此揭示西方媒体在报道中的态度倾向，并分析其中隐含的深层社会历史原因。

*本文系北京市社会科学基金项目“西方媒体中的北京形象：基于语料库的批评话语分析”（15WYC064）的研究成果，也受对外经济贸易大学研究生课程建设项目支持（X17110）。

2. 国内外新闻话语研究现状

新闻话语一直受到语言学界的关注，因为大众传媒能够有力地塑造公众舆论，即从意识形态层面维持并加强社会权利关系（van Dijk 1993：255）。Wodak & Busch（2004）也提出，必须从全球角度研究媒体话语对人们产生和传播信仰和价值观所产生的影响。

语言学界对新闻话语的第一个研究视角是基于语料库的批评话语分析。新闻话语一直是批评话语分析的焦点（Fairclough 1995：3）。批评话语分析理论认为，新闻话语不是对现实的“反映”，而是以新闻生产者期待的方式来“表征”和“建构”现实，并为特定群体的利益服务（Fairclough 1995：3）。批评话语分析主要使用Fairclough的“辩证–关系”方法、Wodak的“话语–历史”方法和van Dijk的“社会–认知”方法（Wodak & Meyer 2009：17-18）。不过，单个文本的话语不具有显著效果，而语料库文本盈千累万，能够为解读话语中的意识形态提供足够的语言证据（Fairclough 1989：160）。因此，Baker *et al.*（2013）提出新闻话语分析的新趋势，即使用语料库辅助的方法进行批评话语分析。他们发现，尽管英国媒体的报道均基于社会事实，但由于不同的新闻媒体代表不同的政治倾向和利益集团，新闻话语在重构社会活动中的权力关系、个人与不同利益群体时，不可避免地受到媒体态度的掣肘。

国内也有一些学者采用基于语料库的批评话语分析方法对新闻话语进行了研究。杨娜、吴鹏（2012）使用1980-2011年*The New York Times*（《纽约时报》）对中国妇女的报道自建语料库，并对语料库的高频词、高频词丛、索引行等进行了分析，发现《纽约时报》通常采用词语“优化”策略、话语风格策略和修辞策略，构建出十分负面、消极的中国妇女形象。刘明（2014）以《中国日报》和《纽约日报》对中美汇率博弈的报道为语料，对新闻话语表征的四种语法结构进行了研究。该研究发现，两份报纸都在新自由主义的意识形态框架下选择表征形式和功能，构建了各自对人民币汇率的立场。兰杰、徐红梅（2014）以英美主流报纸、电台和网站上对中国的报道为研究对象，考察了词汇搭配和词汇–语法结构，发现英美媒体通过使用名物化、被动语态和负面评价词汇等方式，对中国新闻进行了大量负面报道。Yang（2015）收集了《中国日报》BBC news和*The Guardian*（《卫报》）对2008年北京奥运圣火传递的新闻报道，通过分析索引行、搭配等语言特征，发现两国媒体对奥运精神的构建区别很大。中国媒体突出“和为贵”和“人和”的儒家思想，英国媒体则强调“包容”“自由”以及“提倡人权”。作者使用Wodak的“话语–历史”方法框架，从社会和历史背景的角度对两国媒体的不同话语构建进行了分析。

语言学界对新闻语篇的第二个研究视角是结合评价系统和批评话语分析。评价理论是在系统功能语法基础上创建的。该理论认为，评价意义主要通过词汇来表达；它关注说话人/作者如何利用语言资源实现表态功能，探究语篇中所协商的各种态度、所涉及情感的强度，以及表明价值和联盟读者的各种方式（Martin & Rose 2003；Martin & White 2005）。该系统最大的贡献是第一次在语篇层面如此重视人际意义，通过把词汇纳入语义研究范围，克服了纯语法分析的局限性，扩展了人际意义的研究范围，从而丰富了篇章语义学（朱永生、严世清 2001：88）。

评价理论通过词汇层面的研究探讨协商中的各种态度、声音，及其来源和强度，十分符合批评话语分析的批评和评价本质（邱晴 2015：237）。不过，近年来综合使用评价理论和批评话语分析的新闻话语研究仍十分有限。Bednarek & Caple（2010）以*The Sydney Morning Herald*（《悉尼先驱晨报》）的40篇环境报道为语料，使用社会符号学框架和评价系统分析其评价资源，采用批评话语分析和积极话语分析对评价资源的使用进行阐述，以此探究图文结合的新闻文体是否适合进行环境报道。Chiluwa & Ifukor（2015）综合使用评价理论和批评话语分析，对脸书和推特上的社交媒体运动#BringBackOurGirls的话语特征进行了研究。结果发现，抗议活动中表达情感的词汇十分突出，且几乎都是负面评价，符合公众对社会媒体危机话语报道的典型反应。邱晴（2015）以评价理论为主要分析框架，对*The Korea Herald*（《韩国先驱报》）和《中国日报》关于韩国总统朴槿惠就“岁月”号沉船事件向全国人民道歉的英文报道进行了批评话语分析。作者通过对比这两篇新闻中态度资源和级差资源的使用，揭露了语篇隐含的意识形态与社会政治动因。

从以上研究可以看出，基于语料库的批评话语分析方法有助于对新闻语料进行大规模的定量与定性分析；评价理论与批评话语分析在媒体话语研究中也有着天然的联系。目前，借助语料库工具，在批评视野下运用评价理论的新闻话语研究仍十分有限。并且，学界对政治话语的关注较多，但缺少对经济事件相关话语，尤其是中国股市话语的研究。基于以上不足，本研究将使用西方主流媒体对中国股市的报道自建语料库，以Fairclough的三维分析模型为框架，以态度系统为描写层进行分析，力图揭示报道中的态度倾向、态度背后隐含的社会历史原因。

3. 态度系统框架

评价理论根据语义把评价资源分为三个系统：态度（attitude）、介入（engagement）和级差（graduation）。其中，态度系统表达说话人/作者对人或事物的情感和评介，是评价理论的核心，它包含三个子系统：情感、判断和鉴赏。情感系统关注积极或消极情感的表达；判断系统根据伦理、道德和社会规约评论人的行为；鉴赏系统涉及美学评价，依据的是特定领域内评价的方式和标准（Martin & Rose 2003：22-58）。态度意义有正面和负面之分，可表现为铭刻，即通过态度性词汇直接实现；也可表现为引发，即通过隐性手段间接实现。态度系统及其评判说明见表1（Martin & White 2005：48-56）。

表1　态度系统

子系统	类别	定义
情感（affect）	现实型（realis）	快乐/不快乐（un/happiness）
		安全/不安全（in/security）
		满意/不满意（dis/satisfaction）
	非现实型（irrealis）	害怕（fear）
		欲望（desire）
判断（judgment）	社会评判（social esteem）	规范（normality）：How special?
		才能（capacity）：How capable?
		坚韧（tenacity）：How dependable?
	社会约束（social sanction）	诚实（veracity）：How honest?
		正当（propriety）：How far beyond reproach?
鉴赏（appreciation）	反映（reaction）	影响（impact）：Did it grab me?
		质量（quality）：Did I like it?
	构成（composition）	平衡（balance）：Did it hang together?
		细节（complexity）：Was it hard to follow?
	价值（valuation）	Was it worthwhile?

本文将以态度系统作为描写框架，重点考察词汇所表达的显性评价意义。这种意义比较稳定，通常不会随着语境的变化而变化。

4. 研究设计

4.1　语料收集

首先，笔者使用LexisNexis新闻数据库，在英、美八种主要报刊的新闻标题中检索Beijing，收集了它们近5年（2010年10月20日—2015年10月20日）的涉华报道，建成西方媒体“北京”报道语料库。这些报刊包括*The New York Times*（《纽约时报》）、*The Washington Post*（《华盛顿邮报》）、*USA Today*（《今日美国》）、*The Sun*（《太阳报》）、*The Guardian*（《卫报》）、*The Independent*（《独立报》）、*The Times*（《泰晤士报》）和*The Daily Telegraph*（《每日电讯报》）。

其次，笔者使用语料库工具WordSmith Tools，在该语料库中检索stock*，定位其所在索引行（如图1所示）。经过甄别和筛选后，将所有描述中国股市的索引行复制粘贴到文本文档，建成西方媒体“中国股市”报道专用语料库。该语料库包括365条索引行，形符为9,321词，类符为2,088词。语料库的索引行来自2,794个文本，这些文本的容量为115,212词。

N	Concordance	Set	Tag	V
1	Guo Shuqing, who has worked to bring stock market in line with foreign practice,			
2	allows consumer loans to be diverted to stock and property markets; photo (M)			
3	BEIJING INVESTIGATES STOCK MANIPULATION Chinese			
4	since China approved short selling via Stock Connect, due to overly complex			
5	BEIJING COMES UP SHORT WITH STOCK BETS No foreign investor has			
6	issues warning about soaring stock market and tightens margin			
7	left in following week on jitters over stock-market volatility, sign of			
8	$7.3 billion of cash flowed into China's stock and bond markets in three weeks			
9	banks from funneling cash into heated stock market, but caused 7.4% plunge			
10	economy;Â could cut stamp duties on stock sales, but move would hit revenue;			
11	that government sets overall direction of stock market, which has helped boost			
12	Chinese government halts new stock listings and plans massive			

图1　西方媒体“北京”报道语料库中的“股市”索引行

4.2　态度词汇提取

笔者根据表1，在西方媒体“中国股市”报道专用语料库中人工搜索与stock相关的态度词汇。为保证研究结果的可靠性，笔者观察了每个检索词的索引行，根据索引行展示的上下文信息进行判断，确保所提取的态度词汇都是用于描述stock。之后，笔者对这些态度词汇进行手工标注，并统计了不同标注符号的数量（见表2），进一步得到中国股市形象的评价意义分布特征（见表3和表4）。

表2　中国股市形象的态度词汇标注结果

子系统	类别	定义	标注结果
情感（affect）	现实型（realis）	快乐/不快乐（un/happiness）	pain/painful/painfully (4), pessimism (1), sentiment (1), blindly optimistic (1), frustrate (1)
		安全/不安全（in/security）	nervous (2), dangerously/danger (2), warning(s) (2), alarm(ed) (2), panic/panicky/panicking (17), jitter (3), confidence (10), lack of confidence (4), tremors (1), unsettling (1), tension (1), threat (2), not surprised (1), uncertainty (1), risk(y) (9), surprise (6)
		满意/不满意（dis/satisfaction）	complaining (1)
	非现实型（irrealis）	害怕（fear）	fear (s) (21), concern(ed/s) (8), worries/worried (5), terrified (1)
		欲望（desire）	desperate (5), enticing (1)
判断（judgment）	社会评判（social esteem）	规范（normality）	disruption (2), rout (16), disorders (1), convulsion (1), abnormal (1), chaos (4), crisis (24), turmoil (10), jittery (2), volatility/volatile (11), scramble (3), fiasco (1), blowout (2), frenzy/frenzied (5), government-controlled (2), fever (1), gyrations (1), over-heating/over-heated (4), fluctuations (2), restrictions (2), turbulence (1), roller-coaster (1), swoon (2), swings/swung(3), limits (1), meltdown (4)
		才能（capacity）	powerless (1), strong(ly) (3), rapidly (1),surge (3), slow (1), progress (1), gains (5), worst (6), worsens (1), collapse (12), slump (9), tumbled (1), plunge (12), crash (9), strategic (1), futile/futility (2), plummet (3), free-fall (1), bolster (1),soaring (2), flexible (1), incompetent (1), nosedive (2), ham-fasted (1), new (1), inexperienced (1), bust(s) (2), haphazard (1), lack of experience (1), crash(ing) (10), weakness/weakened/weaker (11), short-sighted (1), zoomed up (1), clobbered (1), novice (1), heavy-handed (1), boom (6), artificial (1), confused (1), incoherent (1), stumbled/stumbles (2)
		坚韧（tenacity）	fragility (1)

（待续）

（续表）

子系统	类别	定义	标注结果
判断（judgment）	社会约束（social sanction）	诚实（veracity）	fake (1), true (1), rumor(s) (3), false (1), hearsay (1), speculative/speculation (8), frothy (1)
		正当（propriety）	scapegoating (1), irrational (2), punish (3), crackdown (6), misled (1), exacerbated (1), intervene/intervention(s) (18), defend(ing) (4), plagued (1), malicious (4), imperial (1), dubbed (1), hammers (1), darkens (1), derail(ed) (2), illegal (2), jeopardize (1), to the detriment (1), disrupting (1), accused (1), manipulating/manipulation (6), corruption (1)
鉴赏（appreciation）	反映（reaction）	影响（impact）	lucrative/lure (2), vaingloriously (1), frenzy (3), huge (3), dramatic (3), marginal (1), miracle (1), extraordinary (2), spectacular (1) unprecedented (2)
		质量（quality）	ugliest (1), traumatic (1), dismal (1), catastrophe (1), ironic (1), smooth (1), ridiculous (1), torment (1)
	构成（composition）	平衡（balance）	0
		细节（complexity）	evident (1), visibly (1)
	价值（valuation）	fail/failure (27), not succeeding (2), not paying off (2), unfulfilled (1), loss/losses (21), end badly (2), sink (2), end horribly (1), counter-productive (1)	

表3　中国股市形象的评价意义分布特征

态度	情感					判断					鉴赏			
	害怕	欲望	安全	快乐	满意	规范	才能	坚韧	诚实	正当	影响	质量	细节	价值
正面	0	0	11	0	0	0	25	0	1	0	14	1	2	0
负面	35	6	52	8	1	107	98	1	15	60	5	7	0	59
小计	35	6	63	8	1	107	123	1	16	60	19	8	2	59

表4　中国股市形象的评价意义分布特征汇总

态度	情感	判断	鉴赏	总频次	百分比
正面	11	26	17	54	10.63%
负面	102	281	71	454	89.37%
小计	113	307	88	508	100%

4.3　态度词汇特征

从表4可以看出，在三个态度子系统中，判断类词汇出现最多（307次），情感类词汇（113次）次之，鉴赏类词汇最少（88次）。判断系统是指根据伦理道德标准来评价事物（Martin & White 2005：155），判断类词汇的高频出现表明西方媒体非常关注中国经济的社会道德性。语料库中情感类词汇较多，可见西方媒体在报道中国股市时掺杂着个人情感，不够客观。正如Flower（1991）所说，新闻报道从来不是完全客观和中立的，媒体总是从特定的立场出发来报道新闻事件，并促进和巩固相关立场的再生产。

表4显示，正、负面意义词汇的频数相差很大（正面54次，负面454次），表明西方主流媒体对我国股市的态度非常消极。以下将对情感、判断、鉴赏三个子系统进行详细讨论。

4.3.1　情感系统

表4显示，情感类词汇（113次）几乎都是负面的，正面词汇仅出现11次。这表明西方媒体在报道中国股市时过多地表达个人情感，态度不够客观中立，有失公允。从表3可以看出，正面词汇均属于安全/不安全范畴，包括confidence（10次）和not surprised（1次）（见表2）。尽管confidence重复出现10次，表明西方媒体对中国股市有一定的信心，认为股民对股市行情也抱有一定的信心，但安全/不安全范畴中表达正、负面意义的词汇频数相差仍然很大（正面11次，负面52次）。从dangerously (danger)、panic (panicky)、threat和nervous（见表2）等负面意义词汇可以看出，西方媒体认为中国股市十分危险，令股民惊慌和害怕，不仅伤害中国经济，也威胁世界各国经济。

表3进一步显示，除了安全/不安全范畴，其他四个子范畴都是负面意义词汇。通过这些词汇，西方媒体传达了大量消极意义，如中国股市表现极差，巨大的股市泡沫“诱惑”收入较低的民众入市，使得各方“担忧”，民众一片“抱怨”之声，政府“急需”实施有效举措来挽救股市颓势。如例1显示，中国股市将收入较低的民众“诱惑”入市之后，股市泡沫惊人。由于担心股市泡沫破裂，国家限制股民抛售股票。可见，西方媒体构建了十分消极的中国股市形象。

例1：It had created “an extraordinary stock bubble, by **enticing** people of modest means into the market” and then “became visibly terrified of a burst”, with restrictions imposed on a sell-off of shares.

4.3.2　判断系统

表4显示，判断类词汇（307次）基本上是负面词汇（281次），正面词汇仅出现26次。从表3可以看出，判断系统的正面词汇基本上集中于“才能”子范畴（25次），这些词汇表达了“强大”“快速增长”“进步”等含义（见表2）。仅有1个正面词汇true出现在“诚实”范畴中（见表2），用于表现股市信息真实可靠。其他词汇均为负面词汇，如规范类词汇disorders、turmoil、abnormal、chaos、crisis表达了股市“无序”“不正常”“混乱”“危机”之意；坚韧类词汇只有fragility，表明股市很“脆弱”；诚实类词汇false、hearsay、rumor、fake传达出“虚假”“道听途说”“谣言”等负面信息；正当类词汇irrational、punish、manipulation、illegal等词汇传达出“非理性”“惩罚”“操纵”“违法”等负面意义（见表2）。根据Martin对判断系统的解释，才能与规范类词汇属于社会评判，主要评价人或事物的能力；诚实和正当类属于社会约束，主要评价行为是否真实可信、正当合理。通过使用以上评价词汇，西方媒体刻画了中国股市的不良形象，指出尽管中国股市发展迅速，取得了一定的成绩，但缺乏秩序，暴涨暴跌，十分混乱；中国有关当局的公信力岌岌可危，股民缺乏理性。如例2所示：

例2：Asian stocks fell to three-week lows on Tuesday morning, as a deepening **rout** in Chinese stocks erased risk appetite—sending investors flocking to safe-haven instruments such as government bonds and the Japanese yen.

从例2可以看出，西方媒体认为中国股市混乱加剧，使得亚洲股市跌至3周最低，中国股市不再具有强烈的吸引力，投资者们一窝蜂似地转向更安全的投资手段——买国债、投资日元。

4.3.3　鉴赏系统

表4显示，鉴赏类词汇（88次）中正面意义出现了17次，占19.32%，且集中在“影响”子类（见表3）。“影响”子类中的正面意义词汇（14次）多于负面词汇（5次）。表2显示，正面意义以huge、spectacular、lucrative等为代表，表明中国股市有利可图，发展情况引人注目，潜力巨大。如例3表明，尽管经济学家Nicholls不完全看好中国经济，预测中国GDP增速会由7.8%降至7%，但仍然认为有很多机会找到升值空间很大的低价股票。

例3：Mr. Nicholls is not entirely bullish on the economy as a whole, predicting that GDP growth in China may come down from 7.8 per cent to less than 7 per cent, but says there are plenty of opportunities to find cheaply-priced stocks with **huge** potential.

尽管少数句子传达了部分积极意义，鉴赏系统内的大多数词汇传达的仍然是消极意义（71次，占80.68%）（见表4），尤其是质量和价值两个子范畴（见表3）。表2显示，质量类词汇ugliest、traumatic、catastrophe、dismal、ridiculous传达出“最丑”“灾难性”“荒谬”等负面意义，仅smooth一词表达“平稳”之意。价值类词汇fail、not succeeding、not paying off、counter-productive传达出“失败”“不值得”“适得其反”等负面含义。

总之，从鉴赏类词汇的使用可以看出，西方媒体刻画了中国股市十分消极的形象，认为中国股市虽然有利可图，但股民投机心理严重。一波又一波投资狂潮，导致股市泡沫严重，收益不稳，甚至亏损严重；同时中国政府救市无效，因此股市不值得投资。

5. 社会历史分析

批评话语分析理论认为，新闻话语是一种社会实践话语，它渗透着价值观，会潜在影响读者的价值观（Fowler 1991：26）。Fairclough（1992）也指出，话语隐含权力等因素，对人的身份起着定型作用，并制约着人的社会关系、知识以及思想体系的形成。批评话语分析的主要任务就是挖掘字里行间隐含的意义。笔者将进一步从社会意识形态和历史发展的角度分析西方媒体为何使用大量消极词汇塑造中国股市的形象。

首先，负面词汇的大量使用反映了意识形态对西方媒体的影响。表4显示，西方媒体仅使用了10.63%的积极意义词汇描述中国股市，主要用于认可中国股市的潜力很大。它们对中国股市的评价几乎都是负面的（负面评价词汇454次，占89.37%），且主要集中在判断类。除了“才能”，其余四个子范畴几乎全是负面评价。其中，表示社会约束的“诚实”和“正当”类负面词汇最多，这表明西方媒体对中国股市的社会道德性持否定态度。中国股市在社会道德性上受到的否定评价（如disorders、abnormal、interventions、jitters等）使西方读者不断接收并强化了“中国政府过度干预中国股市，中国股市十分不稳定，缺乏秩序”的负面形象，维护了西方国家对“自由市场”的推崇。实际上，西方媒体一贯宣扬的“自由市场”一次次使人们受创。1929年经济大萧条爆发，这是“自由市场”经济波及面最广、影响深远的市场失灵；1997至1999年，尽管亚洲几个相对开放的股市监管良好，但受到西方基金恶意卖空行为所累，亚洲金融危机爆发，再一次对自由市场提出挑战；2008年，正因为美国的贪婪、无能和监管不力，为全球危机埋下祸根。

此外，西方媒体的负面态度忽视了中国股市的发展阶段，不够客观。中国股市从1989年开始试点，本着“试得好就上、试不好就停”的理念建立，发展历程迄今不过27年。美国股市从1811年纽约证券交易所按照《梧桐树协议》建立开

始，迄今已有200余年。与美国相比，中国股市起步晚了近200年，发展时间很短。此外，美国股市的发展也并非一帆风顺，而是经历了曲折、艰险的过程。从1790年到内战开始，美国股市初步发展；从19世纪末到20世纪初，华尔街加入兼并狂潮，美国股市发展迅速，市场操纵和内幕交易非常严重。直到1929年大萧条以后至1954年，小投资者寻求法律保护，美国股市才开始进入规范发展期（查尔斯吉斯特 2004：18-20）。中国股市的发展时间远远短于美国股市任何一个发展阶段，目前仍处于探索时期。股市体制仍不完善，股民投资心理仍不健全，这是股市发展初期的正常现象。中国政府已出台一系列金融体制改革和股市改革措施，力求促进股市的健康、有序发展。但是，西方媒体忽视了中西方股市发展历史的巨大差异，对其中的任何风吹草动都进行过度解读，不够客观公正。

6. 结语

本文以评价理论中的态度系统为描写框架，研究西方主流媒体近5年对中国股市的报道，并对其态度进行了批评话语分析。研究结果表明，西方主流媒体使用了大量负面评价意义词汇，构建出中国股市的消极形象。通过索引行分析，本研究进一步发现：第一，西方媒体对中国股市的报道不够客观公正，忽视了中国股市起步晚、仍处于探索时期的客观情况。第二，西方媒体特别注重中国股市的社会道德性。在肯定中国股市取得一定进步的同时，大力批判中国经济的道德性。在西方媒体的描述中，中国股民素质不高，为股痴狂；中国股市缺乏秩序，管理混乱。第三，西方主流媒体斥责中国股市的管理方式和行政干预手段，吹捧西方盛行的“自由市场”体制。

本研究使我们清醒地认识到：西方新闻话语反映了其价值观。基于西方媒体对中国股市的负面报道，我们不能对西方媒体改变对中国股市的偏见抱有不切实际的幻想。笔者提出以下建议：第一，中国媒体应加强宣传，揭示西方媒体所谓“客观、公正”报道背后的真面目，使中国民众充分认识到西方媒体“舆论战”对中国的负面影响，提高中国民众尤其是年轻人对西方媒体的判断能力，让更多民众信任中国媒体的立场和观点（冯捷蕴 2013：158）。第二，中国媒体应加强回应负面报道的能力，敢于表达中国的利益诉求，有理有据地进行辩论，维护中国的国际形象。

参考文献

Baker, P., C. Gabrielatos & T. McEnery. 2013. *Discourse Analysis and Media Attitudes: The Representation of Islam in the British Press* [M]. Cambridge: CUP.

Bednarek, M. & H. Caple. 2010. Playing with environmental stories in the news: Good or bad practice [J]. *Discourse and Communication* 4(1): 5-31.

Chiluwa, I. & P. Ifukor. 2015. War against our children: Stance and evaluation in #BringBackOurGirls campaign discourse on Twitter and Facebook [J]. *Discourse & Society* 26(3): 267-296.

Fairclough, N. 1989. *Language and Power* [M]. London: Longman.

Fairclough, N. 1992. *Discourse and Social Change* [M]. Cambridge: Polity Press.

Fairclough, N. 1995. *Media Discourse* [M]. London: Edward Arnold.

Fowler, R. 1991. *Language in the News: Language and Ideology in the Press* [M]. London: Routledge.

Martin, J. & D. Rose. 2003. *Working with Discourse: Meaning Beyond the Clause* [M]. London: Continuum.

Martin, J. R. & P. White. 2005. *The Language of Evaluation: Appraisal in English* [M]. New York: Palgrave Macmillan.

van Dijk, T. 1993. Principles of critical discourse analysis [J]. *Discourse and Society* 4(2): 249-283.

Wodak, R. & B. Busch. 2004. Approaches to media texts [A]. In J. Downing, D. McQuail, P. Schlesinger & E. Wartella. (eds.). *The Sage Handbook of Media Studies* [C]. London: Sage.

Wodak, R. & M. Meyer. 2009. *Methods of Critical Discourse Analysis (2nd Edition)* [M]. London: Sage.

Yang, M. 2015. Olympism and the Beijing Olympic torch relay in the British and Chinese media discourses: A comparative study [J]. *The International Journal of the History of Sport* 32(3): 499-515.

查尔斯吉斯特，2004，《华尔街史》[M]。北京：经济科学出版社。

冯捷蕴，2013，中西媒体话语危机的研究 [J]，《江淮论坛》(5)：140-144。

兰　杰、徐红梅，2014，英美媒体涉疆新闻报道所透视出的价值观 [J]，《新疆大学学报（哲学·人文社会科学版）》(4)：152-156。

刘　明，2014，新闻话语表征的形式、功能和意识形态 [J]，《现代外语》(3)：340-349。

邱　晴，2015，新闻话语的构建、预设及站位——评价系统视角下的新闻报道 [J]，《江西社会科学》(4)：236-241。

杨　娜、吴　鹏，2012，基于语料库的媒介话语分析——以《纽约时报》对华妇女报道为例 [J]，《国际新闻界》(9)：48-58。

朱永生、严世清，2001，《系统功能语言学多维思考》[M]。上海：上海外语教育出版社。

通讯地址： 100029 北京市对外经济贸易大学英语学院

基于语料库的副词“只”与“光”对比研究*

北京语言大学 鲁志杰

提要：本文结合大规模语料库，分析了汉语范围副词“只”与“光”的异同，从句法特点、语义功能和语用衍推等方面揭示了二者的微妙差异。从“只”与“光”的用法出发，观察到二者能否进行替换使用主要受语义韵、单调性衍推、语义相宜性和约束主观小量等因素的制约。

关键词：“只”、“光”、排他、衍推、语料库、语义韵

1. 引言

汉语范围副词的传统研究多从语义分类和界定标准着手（如，张谊生 2000，2004：85；高育花 2001；张亚军 2002；钱兢2005），尹洪波（2011）讨论了否定词“不”与一些范围副词共现时语义变化的三种情况：减量、增量和不变。

近年来，国内针对范围副词“只”和“光”的研究有了较大进展，对于“只”的研究，多从形式语义学视角展开，如蔡维天（2004）、殷何辉（2009，2017）等，此外，卢英顺（1995）从句法和语用角度讨论了“只”与“only”的差异；郭锐（2008，2010）从逻辑语义的角度论述了“只”的语义；徐以中（2003，2010）从语义指向和主观性的角度讨论了“只”的用法；周韧（2015）从现实性和非现实性的角度区分了“只”与“仅”的使用。对副词“光”的研究，学者们主要从历时演变（张琨、许嘉璐 2014）、语义韵（朱希芳2013）、三个平面（马先红、朱希芳 2014）和话题焦点标记（司罗红 2015）等角度进行了讨论。

对范围副词“只”与“光”进行深入对比研究的成果尚不多见。周刚（1999）从语义功能、句法分布、语用选择三方面分析了“光”“仅”“只”的异同，描述了三者的具体用法；王琪（2012）讨论了范围副词“单”“光”“仅”“只”的差异。相对于其他范围副词来说，“只”与“光”的对比研究相对滞后，表现在学界未能以大规模语料库来描述二者的差异。

* 本文初稿曾于首届语言句法分析性与语法参数理论国际研讨会（2018，北京）上宣读，感谢与会各位专家提出的宝贵意见。本文在写作过程中多次得到崔希亮教授和尹洪波副教授的指导，在此致以衷心的感谢！文中谬误概由本人负责。

“只”和“光”是近义词，副词“光”在《现代汉语词典》（2016）和《现代汉语八百词》（吕叔湘 1999）中的解释如下：

只；单。光+谓词性成分。任务这么重，～靠你们两个人恐怕不行 | 产品光好看不行，还得质量好。光+名词性成分。～这点吗？还有吧？ | 不～他一个人，还有别人。

在HSK等级词表中，“只”为甲级词，“光”是乙级词。留学生在初级阶段学习了“只”，进入中级阶段就会遇到“光”，就要对这两个词进行辨析。本文所使用的语料库为“国家语委现代汉语通用平衡语料库”和“北京大学中国语言学研究中心CCL语料库”，使用基于语料库的研究方法，运用Coco Search 1.6展开数据分析，重点讨论副词“只”和“光”的差异，分析二者替换使用的制约条件。

2. 副词“只”和“光”的语料库分析

在“国家语委现代汉语通用平衡语料库词汇频率表”的基础上，生成汉语“只”和“光”的频率表，二者的使用频次分别是6,142和394，可见，副词“只”的使用频率更高。本文以语料库索引为基本依据，结合句法研究，对词的搭配进行分析，概括出范围副词“只”和“光”的句法特点和语义功能。

2.1 “只”和“光”的句法特点

副词“只”与“光”出现的句法位置有：在NP与VP之前、在NP与VP之间，有时会构成“只/光……就”结构，如：

（1）只这一个字就有难以言传的魅力。（《羊城晚报》1984. 12. 29）

（2）光自留地里的黄烟，就卖好几百块。（顾笑言《金不换》）

（3）印欧系古代语言，语法形态非常繁复，我们有时只凭它们的语法构造就可以断定它们是不是亲属语言。（岑麒祥《关于语言亲属关系的问题》）

（4）对此，我们光靠直观思维就不行了。（陆云帆《新闻采访学》）

例句（1）和（2）中的“只”和“光”出现在NP与VP之前，对事物加以限制，由“就”引出VP的内容，这种“只/光+名词性成分+就+VP”的句法结构，表达的含义是从全体的范围列举出部分，以示全体的性质或特征。例（3）和例（4）的“只”和“光”处于NP与VP之间，有时NP可以省略，副词“只”与“光”引导的是介词性结构，VP的内容也由“就”引出，这四条例句中的“只”与“光”可以进行替换使用。通过对语料库的统计，副词“只”与“光”的句法分布如表1所示。

表1　副词“只”与“光”的句法分布

	在NP与VP前	频率%	在NP与VP之间	频率%
只	69	0.0112	6,073	0.9888
光	45	0.1142	349	0.8858

由表1可见，副词“只”与“光”的使用绝大多数会出现于NP与VP之间。在同样可出现的句法结构中，有些句子的“只”不能替换为“光”，如：

（5）a. 据说巴勒斯坦有一族人，办结婚的筵席，只一个菜，是一个烤骆驼。（碧君《岂有斋随笔》）

b. *据说巴勒斯坦有一族人，办结婚的筵席，光一个菜，是一个烤骆驼。

（6）a. 今年他只回过一次家。（《河北日报》1982. 12. 21）

b. *今年他光回过一次家。

（5）和（6）中的“只”分别出现在NP与VP之前和NP与VP之间的位置，且都不能替换为“光”，观察发现，这是由“只”可用来对客观量进行减量的主观评价[1]、限制主观小量[2]造成的。类似的例句还有很多，比较下面两句：

（7）a. 手术组马上把随身携带的气帐篷组装充气，15分钟时间，一座能放置两个手术台的帐篷便搭成投入手术救治，既保暖，又防潮。（《解放军报》1981. 1. 6）

b. 手术组马上把随身携带的气帐篷组装充气，只15分钟时间，一座能放置两个手术台的帐篷便搭成投入手术救治，既保暖，又防潮。（*光）

（8）a. 圩堤龙西段终因经受不住洪水的冲击，出现大面积塌方， 8 米宽的堤面残留 1 米。（《人民日报》1991. 9. 4）

b. 圩堤龙西段终因经受不住洪水的冲击，出现大面积塌方， 8 米宽的堤面只残留 1 米。（*光）

从真值意义上讲，两句的真值没有发生变化，但是在实际表达效果上却是不同的：a句基本是客观陈述，只是表示一定的时间和数量；b句的主观性较强，“只”后的成分可理解为主观小量，还显示出言者认为某个情况的发生比预期的时间短、数量少。

2.2 “只”和“光”的语义功能

“只”与“光”均可以限定范围，但约束对象不同，语义功能上也有异同。根据周刚（1999）对“只”和“光”语义特征的研究，“只”和“光”的原义都表示单一，用于限定范围，具有强烈的排他性，我们认为“只”和“光”的排他性可分为两种，即完全排他和非完全排他。

1）“只”与“光”的完全排他义

从逻辑语义学角度看，郭锐（2008）认为“只”的语义可分析为“选择x，排除非x，x是‘只’约束的焦点成分”，这时的“只”引出的是一个选项的集合{a, x, b, …}，这个义项完全排除a、b等元素。在我们看来，副词“光”也可以作如上的语义分析，表现的语义特征是完全排他义。例如：

（9）有志考大学的学生，许多人只追求分数，什么集体主义精神，人民利益高于一切等等，都置于脑后。（《中国青年报》1981. 10. 17）

（10）光有知识，没有理想和毅力，没有自己动手的习惯和勇于创新的精神，那么，他很可能是个书呆子，而不是实现四个现代化的有用之材。（《中国青年报》1979. 10. 4）

上述例句中“只”可以引出的成分所在集合为{集体主义精神,追求分数,人民利益高于一切，…}，最终选择的是集合中“追求分数”这个元素，完全排除了集合中的其他可选项，由“光”引出的成分所在集合为{有理想和毅力,有知识,有自己动手的习惯,有勇于创新的精神}，例（10）中选择的元素是“有知识”，完全排除了集合中的其他元素，这两个例句中的“追求分数”“有知识”分别是由“只”和“光”约束的对比焦点。

2）“只”与“光”的非完全排他义

非完全排他义多出现于否定句中，此时的否定词“不”“没”与“只”“光”搭配使用，语义关系可表述为“选择x，不完全排除非x”。例如：

（11）在识字教学中，要遵循音形义相结合的规律，不能只重视字音和字形，还必须十分重视字义。（王静《汉字规律与识字教学》）

（12）群众是看我们的行动，而不是光听我们说。（贺中光《双目失明以后》）

这两个例句，虽由于“还”的使用在语义上表现出了轻重之别，但整体表达的语义分别为“在识字教学中，要重视字音、字形、字义”和“群众听我们说的话，也看我们的行动”，“不只/光”与“还/也”组合使用时，可以表达与合取连接词相同的语义真值，因此上述的例句可以作如下变换：

（13）a. 在识字教学中，不能只重视字音，还要重视字形和字义。

b. 在识字教学中，不能只重视字义，还要重视字音和字形。

c. 在识字教学中，不能只重视字义、字音，还要重视字形。

（14）a. 不光风度，他在肚子上、器宇上也像处长。

b. 不光器宇，他在肚子上、风度上也像处长。

c. 不光风度、器宇，他在肚子上也像处长。

“只”所约束的成分可为集合{字音，字形，字义，{字音、字形}，{字音、

字义}，{字形、字义}}[3]中的任一元素，“光”所约束的成分可选自集合{肚子，风度，器宇，{肚子，风度}，{肚子，器宇}，{风度，器宇}}中的任一元素，由于受否定连接词“不”的影响，上述两个例句的变换并不影响命题的真值。由此，当副词“只”和“光”与否定连接词“不”“没”组合时，语句的语义关系为“选择x，不完全排除非x”，“只”和“光”表达一种“包含性排除”，即非完全排他义。

在语料库中，我们对副词“只”与“光”的“完全排他义”和“非完全排他义”这两种语义特征作了统计，副词“只”和“光”的主要语义特征均为完全排他义，如表2所示。

表2　副词“只”与“光”的语义特征统计

	完全排他义	频率%	非完全排他义	频率%
只	6,063	0.9871	79	0.0129
光	364	0.9239	30	0.0761

3. “只”与“光”语用衍推的语料库考察

在自然语言表达中，有些情况不能将“光”替换为“只”，如：

（15）a. 光发言的就有十几位。b. 只发言的就有十几位。

研究发现，这与衍推能力有关。衍推（entailment）是命题间的一种逻辑关系，具有衍推关系的两个命题之间存在着内在的关联性，根据Anderson & Belnap（1962）提出的衍推概念，郭锐（2006）将语义衍推关系分成三种：包容式衍推、释义式衍推和类推式衍推，并将衍推的定义概括如下：

（16）语句p衍推（entail）语句q,当且仅当若p为真，可以由p内在地推导出q为真，记做p⇒q。

3.1　包容式衍推

含有副词“只”和“光”的语句中的对比成分与衍推出的成分具有包含关系或整体部分关系时，是包容式衍推，这时，二者可以进行替换。例如：

（17）a. 张三只吃了苹果。⇒张三只吃了水果。

b. 张三光吃了苹果。⇒张三光吃了水果。

例句中“苹果”与衍推出的“水果”具有真包含关系，“只”与“光”的替换不影响语义表达。

3.2 单调性衍推

郭锐（2008）认为，“只”具有“选择x，排除高于x的成员”的语义，此时集合{a, x, b, …}中的成员具有量级上的区别，呈现一定的衍推序列，此时集合中b以后的成员是排除在外的，这种衍推是单调性衍推（monotone entailment）[4]。

（18）a. 张三只吃了三个苹果。⇒b. 张三没有吃其他的东西。

⇒c. 张三没有吃三个以上的苹果。

上述例句可衍推出两个不同语义的句子，（18b）是前面讨论的“只”的完全排他性语义，在此不做过多的讨论，（18c）中“三个”是“只”约束的焦点成分，“只”仅排除了高于“三个”的成分，并不排除低于“三个”的成分，如“一个、两个”，殷何辉（2009）将“只”的这种用法叫作“未达到更高量级”，副词“光”不能衍推出（18c）的用法。

（19）a. 张三光吃了三个苹果。⇒b. 张三没有吃其他的东西。

⇒*c. 张三没有吃三个以上的苹果。

3.3 非单调性衍推

当含有副词“光”的语句，原句与衍推出的语句中对比成分之间处于同一集合中的平行关系时，不存在释义式衍推中词义上的解释关系，是非单调性衍推(non-monotone entailment)，无法归入到上述分类之中。例如：

（20）a. 光我们班，报名的就有几十人。⇒b. 其他班级也有报名的。

含有副词“光”的语句不属于逻辑衍推的范畴，因此无法归入上述语义衍推的类别中。它是一种建立在情理基础上的语用衍推，常常依赖于特定的语境。这一点与副词“只”有区别：“只”通常表达有量级（scale）上的关系，郭锐（2010：155）指出：“只”的语义表示选择x，排除高于x的成员，是单调向上排除，如：

（21）a. 他光给了十元钱。（除了十元钱没有给其他东西）

b. 他只给了十元钱。（除了十元钱没有给其他东西；没有给更多的钱）

而“光”的衍推是非单调性的，“光”不具有否定高量级的功能。

最常见用于表达非单调性衍推的格式是“光……就”，若用认知语法理论来解释，可视作一种构式，表达“具有平行关系的其他对比成分也含有此性质”的构式义，副词“光”在此构式中表现出了较高的语义强度。例如：

（22）自从朝阳楼饭店“先尝后买，管退管换”的大牌子挂出之后，每天光来吃早点的顾客就有一千五百多人。（*只（《河北日报》1988. 10. 9））

上述例句衍推出的句子语义为“除了早饭，每天来吃午饭和晚饭的人也很

多”，此时不能替换为“只”。

表3　副词“只”与“光”的语用衍推统计

	单调性衍推	频率%	非单调性衍推	频率%
只	369	0.0601	Ø	Ø
光	Ø	Ø	22	0.0558

我们对副词“只”与“光”的衍推用法作了如表3的统计，虽然比例比较低，但仍值得对不同的衍推表现进行分析：当自然话语衍推的句子表现为单调性衍推时，只能使用“只”，不能替换为“光”；衍推出的句子属于非单调性衍推时，须使用副词“光”，不能用“只”。

4. 副词“只”与“光”的语义韵考察

语义韵（semantic prosody）指的是，一个词项在进行组合的过程中，常常会吸引某一类具有相同语义特点的词项，关键词项和与其构成搭配的词项在文本中总是高频共现，这样，关键词项就会十分容易地“感染”上相似或有关的语义特点，整个语境也会弥漫某种语义氛围（Sinclair 1991）。许多学者都对语义韵进行了研究（Louw 1993；Stubbs 1996；卫乃兴2002）。语义韵大体可分为积极（positive）、中性（neutral）和消极（negative）三类。

朱希芳（2013）检索了在《平凡的世界》中“光”的使用情况，结果显示副词性的“光”频数为41，其中有37例为消极语义韵。本文扩大了数据的检索范围，旨在从语料库的角度对现代汉语副词“只”和“光”的语义韵进行考察。语义韵是对语义色彩的描述，可用于表达交际意图和情感态度，受语篇中其他词语的影响，“只”和“光”有时会带有一定的感情色彩，这与二者的语义息息相关。

副词“只”“光”有时与“总是”同义。根据周刚（1999）对“只”“光”的语义进行的研究，“只”和“光”的原义表示单一，而后引申为限定范围，除此之外，二者还可以引申为“老是”“总是”，用以表示频度，如：

（23）a. 他只在食堂吃饭。= 他总在食堂吃饭。

b. 他只看动画片。= 他总是看动画片。

（24）a. 这种布质量不好，光掉色。= 这种布质量不好，总掉色。

b. 她一整天光哭，眼睛都哭肿了。= 她一整天总是哭，眼睛都哭肿了。

图1 “只”“光”的语义：总是

表示“总是”语义的“只”和“光”的语义关系为：“状况p在一定时间范围a至b内间续反复”，状况p是副词“只”“光”的约束成分。例句（23a）和（23b）表达的是“每一次吃饭都是在食堂”和“每一次看电视都是看动画片”，与量化情境变项的“总（是）”基本同义；例句（24a）和（24b）表达的分别是“在一定的情景内反复出现掉色的现象”、“一天内反复重复哭的行为”，“光”量化的是情景单一性的总括信息，因此，与“总是”的语义值基本相同。语义为“总是”的副词“只”所在的句子表达的主要是中性的语义韵，而“光”所在句子的整体语义倾向于消极评价，主观性更强，表达消极语义韵。

副词“只”“光”有“一味”的含义（李宗江、王慧兰 2011），倾向于表达消极语义韵，此时“只”和“光”可以替换使用。例如：

（25）如果不考虑戏的内容和人物形象，只着眼形式，为新而新，为奇而奇，那就有可能走到形式主义的邪路上去。（黄清泽《设计〈红白喜事〉过程的一点体会》）

（26）你不知道他这个性子，就是得管着点儿，不能光由着他。（梁斌《红旗谱》）

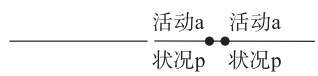


图2 “只”“光”的语义：一味

表示“一味”语义的“只”和“光”表达的语义关系为：“活动a出现时，有且仅有状况p；状况p与活动a相伴而生”，状况p是副词“只”“光”的约束成分，句子的整体语义呈消极评价态度，表达消极语义韵。例（25）表达了言者对戏的看法，对只考虑戏形式的做法评价是消极的，交际的意图是避免走到形式主义的路上，例（26）表达了言者对听者埋怨、不满的情感，言者对“他的性子”和听者对“他”态度的评价是消极的，交际的意图是为听者提建议。

表4 副词“只”与“光”的语义韵统计[5]

	语义	积极	中性	消极
只	总是	37	378	172
	一味	12	147	331
光	总是	1	9	27
	一味	1	13	56

通过语料库，本文对副词“只”与“光”的“总是”“一味”语义韵作了统计（见表4），副词“只”的“总是”语义集中于表达中性和消极的语义韵，其中中性语义韵的比例较高，达到了64%，“一味”语义的语义韵也主要体现为中性和消极色彩的，消极语义韵尤为凸显，比例为68%；副词“光”的“总是”语义多具有中性和消极的语义韵，但消极语义韵的比例更高，占总体数量的73%，“一味”语义的语义韵集中体现为消极语义韵，所占比例高达80%。上述统计数据在一定程度上可解释同为“总是”义，为何有些情况“只”和“光”不能替换使用，观察如下两例：

（23）a. 他只在食堂吃饭。b. 他光在食堂吃饭。

（24）a. 这种布质量不好，光掉色。*b. 这种布质量不好，只掉色。

（23）中的“只”可以替换为“光”，语义基本不变，表达中性的语义韵，而（24）中的“光”却不可以换成“只”。本文认为，这可能与句子整体的语义韵有关，此例句明显地表达了消极语义韵，根据表4，“只”的“总是”语义韵多表现出中性的语义韵，而“光”多为消极的语义韵，因此，在语义相容性上，副词“只”不适合进入该句。

5. 结语

基于语料库的研究不但可以观察到词语的搭配行为，还有利于宏观定性研究。本文运用语料库方法，主要讨论了副词“只”与“光”的异同，分析了二者可以替换使用的条件。

在句法表现上，副词“只”与“光”绝大多数出现于NP与VP之间，有时NP会省略，“只”和“光”引导介词性结构；对事物加以限制，“只”和“光”可出现在NP与VP之前，VP的内容通常由“就”引出，形成“只/光+名词性成分+就+VP”的句法结构，表达“从全体的范围列举出部分，以示全体的性质或特征”的含义。“只”可用来对客观量进行减量的主观评价、限制主观小量，当约束的内容为时间和数量成分时，一般情况下“只”不能替换为“光”。

在语义特征上，“只”和“光”的原义都表示单一，用于限定范围，具有强烈的排他性，体现为完全排他和非完全排他，在这一点上“只”与“光”通常可以替换。副词“只”和“光”的主要语义特征为完全排他义，这时，“只/光”的语义可分析为“选择x，排除非x”，x是“只/光”约束的焦点成分，“只/光”引出的是一个选项的集合{a，x，b，…}，这个义项完全排除a、b等元素；非完全排他义多出现于否定句中，此时的否定词“不”“没”与“只/光”搭配使用，语义可表述为“选择x，不完全排除非x”，x是“只/光”所约束集合{a，x，b，…}中的一个元素，但这一义项并不完全排除a、b等非x的元素。

副词“只”与“光”的衍推用法所占比例较低，当含有副词“只”和“光”的语句可衍推出的句子是包容式衍推时，“只”和“光”可以替换。“只”表达的是单调性衍推，具有“选择x，排除高于x的成员”的语义，集合{a，x，b，…}中的成员具有量级上的区别，呈一定的衍推序列；含有副词“光”的语句是非单调性衍推，最常见的表达形式为“光……就”，表达“具有平行关系的其他对比成分也含有此性质”的语义，在单调性和非单调性衍推用法上，“只”与“光”不能互换使用。

在语义韵的表现上，“只”与“光”存在差异。表达“一味”语义时，二者都倾向体现为消极语义韵；表达“总是”语义时，副词“只”所在语句的语义韵倾向于中性，而“光”大多数为消极语义韵。受语义韵和语义相容性的影响，有些情况下“只”和“光”不能替换使用。

注　释

1. 根据张谊生（2006），主观减量的作用是对客观量进行减量的主观评价。
2. 主观量是含有主观评价意义的量。陈小荷（1994）指出副词“只”是帮助表示主观小量的，一般会重读“只”后边的数量词。
3. 由于只讨论命题的真值，因此在此集合中，并不对元素的顺序加以限制。
4. 单调性衍推，指的是下位、部分概念所在的语句可以衍推上位、整体概念所在的语句，或者上位、整体概念所在的语句能够衍推出下位、部分概念所在的语句，在衍推的过程中存在一种衍推的方向性。
5. 副词“光”的“总是”和“一味”语义表达积极语义韵的例句分别为：
   a. 父亲是标准的干嚎，两只眼睛又枯又呆，光打劈雷不下雨，这种干嚎比湿哭更动人，无数的看殡百姓都被我父亲感动了。
   b. 往日我也养娃，没花一分钱，光凭我这两只奶袋，六个娃长得人高马大。

参考文献

Anderson, A. & N. Belnap. 1962. The pure calculus of entailment [J]. *Journal of Symbolic Logic* 27(1): 19-52.

Louw, B. 1993. Irony in the text or insincerity in the writer? The diagnostic potential of semantic prosodies [A]. In M. Baker, G. Francis & E. Tognini-Bonelli (eds.). *Text and Technology: In Honor of John Sinclair* [C]. Amsterdam: John Benjamins.

Sinclair, J. 1991. *Corpus, Concordance, Collocation* [M]. Oxford: OUP.

Stubbs, M. 1996. *Text and Corpus Analysis: Computer-assisted Studies of Language and Culture* [M]. Oxford: Blackwell.

蔡维天，2004，谈“只”与“连”的形式语义 [J],《中国语文》(2)：99-111。

陈小荷，1994，主观量问题初探——兼谈副词“就”、“才”、“都”[J],《世界汉语教学》(4)：18-24。

高育花，2001，中古汉语副词语义指向分析 [J],《古汉语研究》(2)：41-45。

郭　锐，2006，衍推和否定 [J],《世界汉语教学》(2)：5-19。

郭　锐，2008，语义结构和汉语虚词语义分析 [J],《世界汉语教学》(4)：5-15。

郭　锐，2010，“只”义句和“都”义句的语义等值 [A]，载《语法研究和探索 (十五)》[C]。北京：商务印书馆。

李宗江、王慧兰，2011,《汉语新虚词》[M]。上海：上海教育出版社。

卢英顺，1995，副词“只”和“only”的句法语义和语用比较 [J],《汉语学习》(1)：32-35。

吕叔湘，1999,《现代汉语八百词 (增订本)》[M]。北京：商务印书馆。

马先红、朱希芳，2014，论副词“光”的句法、语义及语用特征 [J],《现代语文》(1)：66-68。

钱　兢，2005，现代汉语范围副词的连用 [J],《汉语学习》(2)：47-50。

司罗红，2015，作为话题焦点标记的“光”[J],《郑州大学学报 (哲学社会科学版)》(1)：127-132。

王　琪，2012，范围副词“单、光、仅、只”的对比研究 [D]。上海师范大学硕士论文。

卫乃兴，2002，语义韵研究的一般方法 [J],《外语教学与研究》(4)：300-307。

徐以中，2003，副词“只”的语义指向及语用歧义探讨 [J],《语文研究》(2)：48-52。

徐以中，2010，“只”与“only”的语义指向及主观性比较研究 [J],《语言教学与研究》(6)：62-69。

殷何辉,2009，焦点敏感算子“只”的量级用法和非量级用法 [J],《语言教学与研究》(1)：49-56。

殷何辉，2017，焦点解释理论对“只”字句语义歧指的解释 [J],《汉语学习》(3)：33-40。

尹洪波，2011，否定词与范围副词共现的语义分析 [J],《汉语学报》(1)：80-85。

张　琨、许嘉璐，2014，限定性范围副词“就”“才”“光”的历时演变研究 [J],《语言文字应用》(4)：141。

张亚军，2002,《副词与限定描状功能》[M]。合肥：安徽教育出版社。

张谊生，2000，现代汉语副词的性质、范围与分类 [J],《语言研究》(1)：51-63。

张谊生，2004,《现代汉语副词探索》[M]。上海：学林出版社。

张谊生，2006，试论主观量标记“没”、“不”、“好”[J],《中国语文》(2)：127-134。

中国社会科学院语言研究所词典编辑室，2016,《现代汉语词典 (第7版)》[Z]。北京：商务印书馆。

周　刚，1999，表示限定的“光”、“仅”、“只”[J],《汉语学习》(1)：13-16。

周　韧，2015，现实性和非现实性范畴下的汉语副词研究 [J],《世界汉语教学》(2)：167-183。

朱希芳，2013,《平凡的世界》中“光”的语义韵和类连接研究 [J],《现代语文》(9)：127-129。

通讯地址： 100083 北京语言大学语言科学院

张爱玲《怨女》自译：出版境遇及译本词汇特征

太原师范学院 史 慧

提要：文学翻译的最终产品是出版物。出版与否虽然不是判断翻译活动成败的唯一标准，但译本内容不可避免会受到出版境遇的影响。在张爱玲小说《怨女》自译中，已知同为英文版的*Pink Tears*和*The Rouge of the North*，只有后者得到了出版机会，但现有文献在双语文本先后顺序记述上存在明显出入。本研究首先在《金锁记》自译框架内商榷《怨女》自译语际方向，然后借助语料库技术分析*The Rouge of the North*文本词汇特征，最后基于张氏英文创作观及出版规律推断*Pink Tears*遭拒原因。

关键词：《怨女》自译、出版、词汇特征、*The Rouge of the North*、*Pink Tears*

1. 引言

在张爱玲小说《怨女》纳入文学自译研究之前，学界对其研究的主流是文学视阈内《金锁记》到《怨女》的重写和再创作过程。通过对比两部小说的叙事，得出了“转译与改写”（周芬伶 2003：296）、“重复、回旋与衍生”（王德威 2004：20-32）等结论。陈吉荣（2009）在其《金锁记》自译个案研究中，做图说明了张氏“历时28年四度改写和自译”《金锁记》，共生成了“五个文本”。其中有三个是《怨女》双语文本。具体而言，“*Pink Tears*在1957年5月被美国出版公司拒绝出版。张爱玲在*Pink Tears*基础上用英文改写成*The Rouge of the North*，1967年由英国出版公司出版，并于次年将其译成中文版《怨女》，由台北皇冠出版社出版。”至此，《怨女》自译属性明确，且语际方向被描述为英译汉。然而，据陈亚明（2011：66-67）《张爱玲译事年表》一文梳理：“1956年10月张爱玲完成英文长篇小说《粉泪》。1957年《粉泪》未被Charles Scribner’s Sons公司接受出版……1966年，小说《怨女》中文版在香港《星岛晚报》连载。4月，《怨女》单行本由台湾皇冠出版社出版。1967年，长篇小说*The Rouge of the North*《北地胭脂》（《怨女》英文版）由英国凯赛尔公司（Cassell & Company）出版。”显然，两位学者关于《怨女》中文版出版时间和双语文本先后顺序的记述存在出入，而文学自译研究一旦忽略双语文本出版细节，极有可能走入关于自译语际方向等问题的认知误区，进而影响自译研究结论的生成。鉴于此，本研究首先商榷《怨女》自译语际

方向，然后借助语料库技术，分析《怨女》英文自译本*The Rouge of the North*基于《金锁记》的改写方向及其与同时代英美非翻译小说的区别。

2.《怨女》双语文本创作与出版

2.1 与《金锁记》自译的嵌入关系

《怨女》自译活动嵌套于《金锁记》自译格局之内，不宜与之割裂而谈。二者相嵌的契机是张氏在1956年完成了将中文小说《金锁记》改写为英文小说的创作，题为*Pink Tears*[1]。

据文学史料，1943年《金锁记》发表于杂志《万象》，被当时笔名迅雨的傅雷先生盛赞为“中国文坛上最美的收获之一”。小说很快受到热捧。志在“要比林语堂还有出风头”（刘绍铭 2013：31）的张氏迫不及待地规划并创作了*Pink Tears*等一系列英文小说。然而，改写自《金锁记》的*Pink Tears*（即后来在英国出版的*The Rouge of the North*和在台北出版的《怨女》）几经修改却出版无果（张爱玲等 2011：31-51）。尤其是曾在1955年出版张氏*The Rice-Sprout Song*（《秧歌》英文版）的纽约司克利卜纳（Charles Scribner’s Sons）公司在1957年对其新作毅然拒稿。此举对张氏可谓当头棒喝。所幸，张氏转而先将*Pink Tears*汉译为《怨女》，再将《怨女》英译为*The Rouge of the North*。截至1967年，《怨女》自译完成。其中，中文版和译界公认的英文版*The Rouge of the North*均得以出版，*Pink Tears*止于遗珠之恨。1967年8月张氏自译《金锁记》，题为*The Golden Cangue*，1971年由哥伦比亚大学出版社出版（陈亚明 2011：68）。

我们发现，在始于《金锁记》、止于*The Golden Cangue*、由2个中文本和3个英文本构成的自译链条上，《金锁记》与*Pink Tears*呈跨语种改写关系；《金锁记》与*The Golden Cangue*、*Pink Tears*与《怨女》、《怨女》与*The Rouge of the North*等3对文本分别呈自译关系。从发表结果上看，除*Pink Tears*之外，其余4个文本皆在相应语种语境发表。

2.2 语际方向商榷

王德威（2004：18）在《落地的麦子不死：张爱玲与“张派”传人》一书注释部分明确列出：“《怨女》英文原稿名为*Pink Tears*，于1958年即可能做出，为张于MacDowell写作营的作品。见司马新《张爱玲与赖雅》（台北：大地出版社，1996），81，98，126，138页。”鉴于该书的文学属性，《怨女》的自译属性、语际方向以及另一英文版的存在并未提及。陈吉荣（2009：68-89）梳理了“存在互文关系”的《金锁记》五个文本的顺序：*Pink Tears*为第二文本，*The Rouge of*

*the North*为第三文本，《怨女》则为第四文本。文本间关系被界定为：一三文本之间、三四文本之间都是“严格意义上的翻译关系”，一二文本之间、二三文本之间是“改写”关系。《金锁记》双语文本属“延时自译”，《怨女》双语文本则是“即时自译”。可见，该研究用“互文”定义《怨女》自译与《金锁记》自译的相关关系；认为*Pink Tears*之后，*The Rouge of the North*先于《怨女》而存在，因此《怨女》自译的语际方向是由英语而汉语的。此处关于语际方向的结论明显与《张爱玲译事年表》中信息相悖，亟待考证。

2013年由新星出版社出版的《张爱玲的文学世界》一书第45页上，“张爱玲毕生之友、翻译家宋淇与邝文美之子、现为张爱玲文学遗产执行人，致力于张爱玲文学遗产的保护与整理”的宋以朗先生对《怨女》双语文本产生的孰先孰后做出了如下还原：

> “……我是从一个英语读者的角度去看张爱玲的英语长篇小说，一共有六本，可以分前期和后期，分水岭是一九五五年张爱玲离开香港到了美国。前期的英文长篇小说是这三本——《秧歌》、《赤地之恋》和《怨女》，其中《秧歌》是先写成英文*The Rice-Sprout Song*，后翻译成中文；第二本是中文《赤地之恋》在先，英文*The Naked Earth*在后；第三本是先写了英文的*Pink Tears*（《胭脂泪》，又译《粉泪》），然后翻译成中文的《怨女》，再翻译成英文的*The Rouge of the North*（《北地胭脂》）。”

很明显，宋以朗先生的文化身份和资料优势对于《怨女》自译研究的意义至少有以下三点：首先是帮助确认了《怨女》双语文本自译关系属性；其次是明确还原了双语文本的创作顺序，即*Pink Tears*——《怨女》——*The Rouge of the North*、英文——中文——英文；再次是提醒我们《怨女》先后经历了两次语际方向截然相反的自译，产生了3个双语文本，其中英文版有2个（图1）。

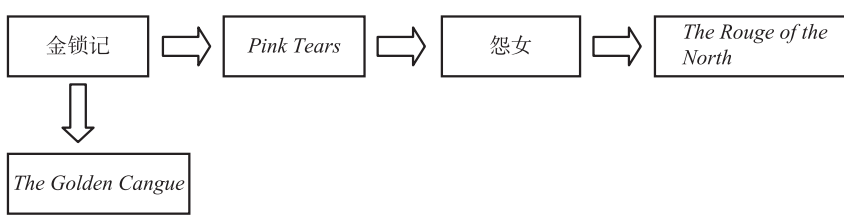


图1 《金锁记》双语文本与《怨女》双语文本关系图

鉴于译本出版是较为复杂漫长的过程，文本的出版时间便不宜与创作时间画上等号。读者和研究者一旦对文本创作顺序产生“误会”，对自译活动的认知便会产生偏差。显然，在有文献记载的前提下，自译语际方向的认定建立在创作时间基础之上较为稳妥。

3.《怨女》英文自译本分析

《怨女》英文版有二，出版命运截然相反。自译者张爱玲之所以对《怨女》所负载的中国故事两度自译，不排除受制于强烈的出版动机。那么，张氏如何在迎合英美出版环境和保留中国故事精髓之间做出取舍、达成妥协，便是《怨女》自译研究中最有必要关注的焦点。我们推测，首先，*Pink Tears* 文本特征与出版遭拒之间很可能存在某种必然的联系；其次，其与 *The Rouge of the North* 间的文本差异导致出版境遇差异；再有，*Pink Tears* 文本缺失的现实决定了文本差异研究只得从 *The Rouge of the North* 与 *The Golden Cangue* 文本差异入手，推定微观改写方向，从 *The Rouge of the North* 与同时代英美非翻译小说的区别，分析其宏观出版环境。

3.1 基于《金锁记》的改写方向

整体上，改写引发了篇幅骤增：《金锁记》到 *The Golden Cangue*，44 页变成 66 页；《怨女》为 345 页，*The Rouge of the North* 长达 185 页。（宋以朗、符立中 2013：55）细节上，*The Golden Cangue* 是《金锁记》的“直译本”；*Pink Tears* 是“直接”改写本，*The Rouge of the North* 则是“间接”改写本。因语内转换不及语际转化复杂，两改写本间差异宜最小化（陈吉荣 2009：69）。下文拟对比其独特词，即“在一个文本中词频达到一定水准而在另一个类似文本或者其他多个类似文本中词频为零的词语”（冯庆华 2008：269），以便直观揭示改写方向。

我们使用 Concordance 320 软件。首先将 *The Golden Cangue*（简称 GC）与 *The Rouge of the North*（简称 RN）两个英文本拷贝至同一个 TXT 文件，在文本分界处键入一串字符，如 20 个 0，作为分界标记。然后点击 Make full concordance from files 图标，添加该 TXT 文件，右键选择 Sort by occurrence。调换两个英文文本的顺序，依照同一方法再次操作，得到两张独特词表。表中从 0 字符串下方的词汇起即为第二个文本的独特词。按照词性分拣两文本独特词的前 100 个。

（1）名词

GC 独特名词 22 个：cangue（3 次）、moons（1 次）、moonlight（4 次）、teardrop（2 次）、barefoot（1 次）、color（5 次）、stain（2 次）、pallet（3 次）、crevices（2 次）、peasants（1 次）、quarters（2 次）、attraction（1 次）、breed（2 次）、season（1 次）、customers（1 次）、TO-YüN（1 次）、FENG-HSIAO（19 次）、FENG-HSIAO’s（1 次）、CHIANGS（14 次）、SHUANG（25 次）、TS’AO（10 次）、CH’I-CH’IAO（126 次）。

据上可得 RN 就《金锁记》进行的删除痕迹，大体分两类。第一类是与《金》题目及开篇段落删除直接相关的内容缺失：（1）RN 文本无“锁（cangue）”，张氏改写过程中从题目到内容弃用“金锁”意象，恰证改动决心；（2）RN 删去了《金》开篇段落，最典型的是舍弃了为小说定下悲凉叙事基调的 6 处“月亮（moons）”：“三十年前的上海，一个有月亮的晚上……我们也许没赶上看见三十

年前的月亮。年轻的人想着三十年前的月亮该是铜钱大的一个红黄的湿晕，像朵云轩信笺上落了一滴泪珠，陈旧而迷糊。老年人回忆中的三十年前的月亮是欢愉的，比眼前的月亮大、圆、白；然而隔着三十年的辛苦路望回看，再好的月亮也不免带点凄凉。”放弃了对月亮形状、颜色、情绪等集中刻画，转而使用多元意象，如人、城市、电灯、天色、房子等：“上海那时候睡得早，尤其是城里，还没有装电灯。夏夜八点钟左右，黄昏刚澄淀下来，天上反而亮了，碧蓝的天，下面房子墨黑，是沉淀物，人声嗡嗡也跟着低了下去。”通过昼夜、天地、明暗、喧闹与静谧等时间、空间、视觉、听觉等多重对比，将读者的观感引向暗夜中“沉淀”出的静谧，引人屏息聆听；于是，该段落中的“朵云轩（To-yün Hsüan）”、“泪珠（teardrop）”、“湿晕（stain）”也随之弃用；第二类是因人物删除导致的内容缺失：RN删去了《金》姜公馆三奶奶的仆人凤箫（Feng-hsiao）和二奶奶的仆人小双（Little Shuang）、女主人公曹七巧的儿媳芝寿（Chih-shou）、大奶奶玳珍（Tai-chen）、姜家三爷季泽（Chi-tse）、七巧女儿长白的订婚对象童世舫（Shih-fang）、七巧的哥哥曹大年及其子曹春熹等人物，导致用以烘托凤箫和芝寿心境悲凉的“月光（moonlight）”、仆人小双“光脚（barefoot）”走到窗边看月亮、凤箫和小双说悄悄话的下房（servants’ quarters）、地铺（pallet）、房间的窗户眼儿（crevices）等空间细节、小双说仆人的三季衣裳（Of my wardrobe for the **season**）都像庄稼人（peasants）、小双揭穿主人七巧曾是麻油店的活招牌（she was the big **attraction** at the sesame oil shop）、贬低其出身低微（Dragons **breed** dragons, phoenixes **breed** phoenixes）且见多识广（dealing with all kinds of **customers**）、人物衣服或生命的颜色（color）、季泽、童世舫等人的脸色（color）、七巧在暗夜里感怀三十年悲惨人生落在枕上的泪珠（teardrop）、教训女儿长安时房间漏风的窗户缝（crevices）、分家后七巧家里下人潘妈所在的下房（the amahs’ **quarters**）等细节一应删除。

RN析出35个独特名词：o’clock（2次）、evening（3次）、summer（6次）、clearing（1次）、city（13次）、shops（4次）、square（7次）、shopfronts（1次）、boards（6次）、accompaniment（1次）、palm-leaf（5次）、chinks（1次）、shake（3次）、toes（2次）、bangs（6次）、brows（4次）、beauty（6次）、buttocks（1次）、spectators（1次）、shutter（1次）、peep-hole（4次）、peach（4次）、blossom（1次）、shavings（1次）、hole（7次）、wisp（1次）、sediments（1次）、whiff（2次）、range（1次）、spindle（1次）、system（6次）、sickness（3次）、customer（1次）、arch（2次）、coolness（1次）。据上可得RN新增内容。值得注意的是，仅开篇位置就涵盖了RN全文约91%的独特名词，多达32个。

情节上，与《金》开篇烘托30年前后物是人非的月亮描写及两个仆人在月色中交代姜公馆主仆关系不同的是，《怨》开篇是城市的黄昏中一名形骸放浪的痴汉去麻油铺撩拨女主人公银娣的情节。其中引入了小时（o’clock）和夜晚（evening）等时间概念、上海（Shanghai）、城市（city）、商店（shops）、街

道（street）等空间概念；感官上，《金》侧重的是月光、地铺、窗户缝儿等视觉意象，借仆人凤箫和小双之嘴交代女主人公曹七巧成为姜公馆二奶奶的原因，为七巧一生无爱、孤独老去的命运埋下伏笔。而《怨》从亮着的天（sky clearing）、作为沉淀物（sediments）的墨黑的房子、京剧间歇人声模仿的伴奏（accompaniment）、痴汉手中的芭蕉扇（palm-leaf fan）、麻油店门洞（peep-hole）等视觉、听觉双重角度激发读者的想象力，通过安静的城市街道上痴汉前往麻油铺的疯癫样态暗示女主人公银娣引人垂涎的美貌及出身低微的事实；从线索人物的性别及数量来看，《金》开篇是女性群像，而《怨》开篇是就单个男性的特写。以上区别佐证了张氏在改写中规避重复的创作意图。

据独特名词语境推测改写细节，我们发现在《金》相对粗线条的叙事基础上，张氏自《怨》开篇起就着力完成以下任务：1）明确设定燥热夏季（summer）为小说时间背景，故多数人物手持芭蕉扇（palm-leaf fan）、喝夏日饮料（summer drinks）；2）加入女主人公银娣家的麻油铺内外陈设，如店门（shopfronts）、门板（boards）、门板夹缝（chinks between the boards）、门洞口（peep-hole）、对面药店家门板上开着的方洞（square on the door/peep-hole）；3）增加银娣的外貌描写，如刨花头油（the pungent sticky juice of wood shavings that women put on their hair）、薄薄红嘴（bright pink and sculpted lips）、短脸配着长颈项与削肩（a short face on top of the long neck and sloping shoulders）、剪成人字式的刘海（bangs cut into a pointed arch）、眉心梭形的揪痧痕迹（a small purplish red mark stood like a spindle between the brows）等；4）添加中国特色民俗文化元素，如痴汉哼唱的京剧（Beijing Opera）、模仿的胡琴伴奏（accompaniment）及《斩黄袍》选段的唱词“孤王酒醉桃花宫，韩素梅生来好貌容（I, the king, drunk in the Peach Blossom Palace, With Han Su-ngo of beauty matchless.）”、踱的方步（square step, walking with feet wide apart, toes pointing outwards）等京剧元素、银娣揪痧去热（pinch the heat sickness out of the system）的中医疗法、银娣大婚时前来贺喜的周遭店铺（shops）、洞房床板（bed boards）、银娣头上红色方形喜帕（the square of scarlet cloth flung over the head）、哥哥炳发给不出的满月礼（the Full Month gift）、银娣给儿子玉熹纳的小妾冬梅方形的象征多子的臀部（square behind）、姚家三房因珠花失窃需寻找嫌疑人而发起的圆光法事（hire a round lighter）等婚俗及迷信活动；姚家老宅天井下方的青石板地（stone-paved square as into a well）等建筑元素等。

（2）动词

从数量上看，GC析出14个独特动词，RN则多达29个，是GC的2.07倍。根据阅读经验，动词使用越多，创作者将故事情节细化、明晰化的程度越高。使用AntWordProfiler1.3.1w软件考察动词难度：GC独特动词中高达64.3%属于较复杂、较正式的非基本词汇，如：blurred、tinged、discards、economize、musing、tittered、sneered、rime、sneezed等；RN独特动词有62.1%属英语基本词汇，仅

有37.9%的独特动词属于复杂度和正式度较高的非基本词汇，如slither、flapped、retraced、snapped、pounding、grumbling、clatter、sculpted、looming等。可见，张氏在改写过程中努力提高《怨女》情节的易读性。

（3）形容词

GC独特形容词16个，其中与开篇月夜、月色等悲凉气氛相关的有Moonlit、reddish-yellow、whiter、apt；与凤箫和小双对话场景及内容相关的有blue-white、oil-green、criss-crossed、Korean、personal、half-worn、quilted、cuddlesome，涉及月下人物肤色、裤管颜色、毯子样貌、众仆人睡姿、丝绸产地和质地、仆人小双与七巧的主仆绑定关系及小双的可爱样态等内容；与女主人公七巧及丈夫有关的有passable、titled、crippled、vulgar，涉及七巧的嫁妆寒酸、出身低微、语言粗俗、丈夫残疾等细节。

RN弃用上述形容词，从开篇处另行使用独特形容词共12个，其中与上海的夜晚描写有关的有blue-green、pebble-paved、electric、uncommon、lamp-lit，修饰的对象为天色、道路、电灯、麻油铺门洞等；与痴汉情节相关的有musical、matchless，涉及哼唱的胡琴伴奏和京剧唱词；与银娣的外貌相关的是winglike、sloping、pungent、unreal、neat，勾勒银娣的发型、头油、肩膀、面庞等细节。从获得以上形容词的感觉器官来看，GC侧重表颜色、形状、质地等内容的视觉类；RN则除视觉类以外，还纳入了听觉和味觉类形容词，可见张氏改写过程中丰富感官维度的努力痕迹。

（4）副词

在位列前100的独特词汇中，GC有softly和faithfully两个副词，RN文本有happily、busily、outwards等3个副词。通过Concordance 320提取副词语境，发现从数量和质量上并不能有效区别两个英文文本。

（5）其他

RN弃用了代词ours和序数词seventh，据语境发现，弃用是《怨》中删去凤箫、七巧等人物所致，相应地，凤箫称呼自己主人、七巧嫂子称呼七巧时惯用“我们家（奶奶）”、“咱们家姑奶奶”等表达以及七巧因生于七月而得名等信息便随之弃用。

以上依据词类分拣出的改写细节一方面印证了张氏“非常谨严”的写作态度：确保面貌体型都有明确的轮廓纹，否则会自觉心虚（张爱玲等 2011：9）。另一方面也提供了张氏改写的方向线索：开篇方式改换、线索人物调整，主人公样貌及出身细节化、中国特色民俗内容凸显。

3.2 与英美原创小说的区别

当*Pink Tears*铩羽而归，*The Rouge of the North*欲得出版机会必然要求张氏以

迎合英美文学场为前提，同时确保中国故事独具特色，不致淹没于英美原创小说的洪流之中。

AntConc 3.3.5w软件Keyword List功能生成的“关键词表”可以完成区辨文本内容这一任务。其工作原理是通过将目标语料库和参照语料库两库的词频表进行对比，统计生成其中一个词频表中明显高于另一个词频表的那部分词汇项目表。通过呈现主题性（keyness）值，即“跨库频率显著性”，来凸显目标语料库的主题或内容特色。

使用AntConc 3.3.5w，以RN为目标语料库、6部英美小说[2]为参照语料库提取“关键词”。经观察，发现RN与6部非翻译英美小说的词表降序排列相对吻合。此现象一定程度上说明张氏英文小说的出版境遇与英文写作水平无关，与创作内容更为相关。

总体上看，在主题性降序排列及负主题性降序排列指标下，我们观察位列前100的数据。首先，RN主题性值最高的是非人物姓名类完整词汇mistress和third，该值分别为1,277.993和909.370，排在第三位的才是人物姓名类词汇yindi（银娣）。RN前100关键词分为三类：其中，非姓名类完整词汇高达82个；人物姓名词汇或词素有15个；非人物姓名类词素3个，列51、78等位置。按照词性统计非人物姓名类完整词汇，发现8类：名词、形容词、序数词、动词、副词、代词、冠词、介词等。RN文本44个名词按照内容大致分为6类：1）人物关系类：mistress、sister、brother、master、slave、amah(s)、concubine、couple、bride、wife、relatives、grandmother、family、son、relative；2）生活物品类：fan、bamboo、gold、gown(s)、silk、oil、mahjong、pigtail、lamp、couch、pot、jacket、bean、dung、money、mandarin；3）行业指示类：singsong；4）空间类：Shanghai、Peking、shop、alley、row、courtyard、pharmacy；5）时间类：year、today；6）交通工具类：ricksha(s)。RN文本中析出的代词呈5倍数值，有it、everybody、they、somebody、she等5个。

同时，据负主题性降序排列的关键词列表，反向考察RN能够揭示同时代主流英美原创小说的主题内容。首先，6部英美小说关键词第一位是代词，而RN代词I的使用以888.47的主题性远低于前者，即弃用了6部英美小说偏好的第一人称单数“我”叙事。观察6部英美小说偏好使用的词汇，析出的词性虽然包括代词、名词、连词、情态动词、感叹词、介词、动词、助动词、形容词、副词和冠词等多达11种，但明显缺失RN文本中具有明显词频优势的数词。具体而言，以下4种词性差异较为显著：

（1）代词

6部英美小说偏好代词的使用：第一人称单数I、me、my、myself；第二人称you、your；第三人称单数he、his、its以及关系代词which、that、whom。RN析

出的张爱玲偏好使用的it、everybody、they、somebody、she，可见张氏叙事人称相对单一，只集中在第三人称（it、they、she）和泛指类人称（everybody、somebody）上，且第三人称倾向于女性视角。

（2）名词

6部英美小说偏好以下类别名词：1）人物称呼类：Mr.、Mrs.、sir、miss、gentleman；2）主观情绪情感类：mind、pleasure、heart、soul、fear、feelings、pain、happiness、words、cause；3）政治及信仰类：state、god；4）日常生活意象类：letter、friend、dress、horse；5）时间空间类evening、town；6）人物类：friend。相比之下，RN中张氏则侧重：1）家庭人物关系类：mistress、master、wife、relatives、amah(s)、people、nana、concubine(s)、son(s)、grandmother、bride、brother、family、slave、sister、servants、couple；2）娱乐消遣类：singsong、opium、mahjong、shop；3）时间空间类：year、sun、pharmacy、today、courtyard；4）生活用品类：rice、silk、lamp、pigtail、couch、oil、row、rickshaw(s)、gown(s)、cloth、bowl、jacket、fan、bamboo、bean、dung、money、pot；5）地名类：Shanghai、Peking。从小说对社会的关注视角来看，6部英美小说着眼于政治和国家的“大社会”，张氏偏好家庭内部的“小社会”，至多是在“小社会”里纳入对上海、北京两地的局部观察和描绘，创作题材更重“男女间的小事情”，因张氏认为人在恋爱时比战争或革命时更素朴、更放肆，最能流露真性（张爱玲等 2011：9）。从人际关系的界定看，6部英美小说倾向于按照性别对人物划分和称呼，“朋友”一词的大量使用更是社会追求“人人生而平等”的写照，而张爱玲则选择家庭内部长幼、次第、尊卑为划分标准，再现封建社会旧式家族的人际风貌。英美原创小说关注人物的心理活动描绘，思想、情感、幸福、痛苦都在刻画内容之列，张爱玲则更倾向于借由小说实现中国特色娱乐消遣、生活用品的跨文化跨语际播迁。

（3）连词

6部英美小说在连词的使用上包含了并列、转折、递进、选择、因果等类别；而RN关键词中连词缺失，很大程度上折射出了汉语母语写作中鲜少利用连词交代上下文逻辑关系这一习惯对英文行文的影响。连词“不显著”想必会给读者造成行文逻辑稍嫌松散的阅读印象，极易加重理解负担。

（4）动词

6部英美小说的偏好动词主要有3类：1）判断类：am、are、been；2）思想情感类：thought、love(d)、cried、wish、hope、think、exclaimed、supposed、hate、thank；3）其他类：answered、began、continued、read、try、done、received、return、observe、saw。RN文本关键动词只有got、getting、padded等3个，从数量和种类上都不能和英美小说动词的丰富程度和表现力媲美。根据常识，英语每一个句子都有至少一个动词。可想而知，在读者期待亮彩、获得重要信息的绝大多数动词位置，张

爱玲的英文作品都略嫌平淡和单一地“一笔带过”。虽然我们不能武断地对此现象进行归因分析，然而有一点却毋庸置疑，那就是张氏更擅长使用名词为作品的表现力增色。

4.《怨女》文本外因素对出版结果的影响

4.1 创作内因素：英语文学创作观

整体上，张氏英文创作始于对文学的热衷、超越林语堂、韩素音的志愿、经济压力以及把中国文化介绍到西方的强烈愿望。在大多数作品中，张氏喜欢将自己抽离，从旁观察，故角度别开生面。一来本性使然，二来可能跟她因为“市场需要”而刻意“用洋人眼光来看中国”的习惯有关，于是“深入本质”“尝试解释”，以期西方人明白（宋以朗 2015：275-276）。当西方出版商因故事复杂而对其拒稿时，张氏的习惯做法是不放弃甚至是重写。张氏的重写是重新写，而非重复写，是“逐次渐进而非原地踏步”，其思想情感上每一次重写都透入深一层灵魂探索（宋以朗、符立中 2013：7-9）。面对出版商对其“取材狭窄、单一”的指控，张氏曾在《自白》中坦言自己的写作受中国古典文学和新文学影响，语言隔阂和文学传统便成为西方读者接受其故事的双重障碍。”（同上：19）而关于创作内容，张氏直言“只写最熟悉的、坚持最熟悉的”。据此，我们有理由站在创作者张氏的角度为其辩护，所谓“创作视野狭窄”是其为刻画普通人面对战争时人性的反常和扭曲做出的主动选择，即便是改写也要不断求变，因为不认同美国小说家马昆德（John P. Marquand）自传式风格（张爱玲等 2011：51-52）。

面对赴美后每一部英文小说从写作到出版必经的甘苦历程，张氏也展现了向英语文学及文化圈讲好“中国故事”的坚定决心和不灭热情：首部英文小说《秧歌》创作完成后等待美国经理人的回音、求“牙牌签”问卜运气、出版后虽得到《纽约时报》《星期六文学评论》《先驱论坛报》《时代》等主流报纸杂志好评，却最终落入美国小说界规则怪圈——非畅销书即遭遇书商拒绝再版（同上：22-27）；其后的《赤地之恋》书成后因创作大纲由他人拟定（commissioned）、印刷水平低、宣传不充分等原因无人问津（同上：28）。1978年7月19日宋淇致信张氏帮助其分析销路不佳主要是因为“事过境迁，读者已不感觉切身之痛，提起韩战，美国、中国年轻人知都不知道。”（宋以朗 2015：249）

4.2 创作外因素：出版规律

除却*Pink Tears*等出版遭拒作品，我们更有必要总结成功发表的张氏英文作品并发现出版规律。按照独立创作时间，以开启过出版流程且是自译活动的源语文本或译入语文本为基准，梳理张氏英文小说出版史实，可得表1。

表1　张爱玲英文小说出版情况统计

序	小说英文名	对应中文小说	类型	方向	创作完成年份	出版与否	英文版出版年份/国家或地区/机构
1	*Naked Earth*	《赤地之恋》	长篇	汉译英	1954	是	1954/中国香港/联合出版社（Union Press）
2	*The Rice-Sprout Song*	《秧歌》	长篇	英译汉	1955	是	1955/美国纽约/司克利卜纳公司(Charles Scribner's Sons)
3	*Pink Tears*	/	长篇	/	1956	否	1957/为上所拒
4	*The Stale Mates*	《五四遗事》	短篇	英译汉	1956	是	1956/美国/*The Reporter*（《记者》/《通讯者》）双周刊
5	*Little Finger Up*	《等》	短篇	汉译英	1961	是	1961/收入吴鲁芹（Lucian Wu）编写的《中国当代作家小说12篇》（*New Chinese Stories: Twelve Short Stories by Contemporary Chinese Writers*）/中国台北/Heritage Press
6	*Shame, Amah!*	《桂花蒸·阿小悲秋》	短篇	汉译英	1962	是	1990/收入Ann C. Carter 和Sung-sheng Yvonne Chang合编的《雨后春笋：台湾女作家当代小说》（*Bamboo Shoots After Rain: Contemporary Stories by Women Writers of Taiwan*）/美国纽约/The Feminist Press
7	*The Rouge of the North*	《怨女》	长篇	汉译英	1966	是	1967/英国/凯赛尔公司（Cassell & Company）
8	*The Golden Cangue*	《金锁记》	中篇	汉译英	1967	是	1971/收入夏志清（C.T. Hsia）编《中国现代中短篇小说选》（*Modern Chinese Stories and Novellas 1919-1949*）/美国纽约/哥伦比亚大学出版社（Columbia University Press）（Hsia & Lau 1971）

从出版现象上看，张氏实际创作8部小说，成功发表7部，其中就篇幅而论，长篇3部、中篇1部、短篇3部；从自译方向上看，英译汉2部，汉译英5部；英

文本创作时间1954年到1967年，始于张爱玲离港赴美前后；发表时间几乎同步，略有延迟，延迟时长短则1年，长则8年（如*Shame, Amah!*）；总体上看，长篇小说的出版机构多为英美，作家因无固定合作伙伴，所以除了努力呈现叙事及语言功力以外，只能焦灼等待发表时机，成败概率各半：如*Naked Earth*书成之后，美国出版商“果然没有兴趣”（张爱玲等 2011：28）；*The Rice-Sprout Song*和*The Rouge of the North*在“对以中国为题材的小说有不同审美”的境外文学场登陆成功，想必与美国经理人玛莉·勒德尔（Marie Rodell）女士的专业运作分不开（宋以朗 2015：282），因封面上赫然写着“A Novel of China Today”的字样。而中短篇小说除*The Stale Mates*[3]以外，发表多仰仗华人作家宋淇、邝文美夫妇及夏济安、夏志清、聂华苓、前美国新闻处文化部（USIS）主任麦卡锡[4]、美国小说家马昆德[5]等举荐提携[6]。

从出版失败的结果上看，张氏总体上将之归咎于“语言障碍外的障碍”，认为“西方对中国和中国人有想法。写出来的东西不符合外国人的想法，他们就不满意”；接受的人多半是“中国通”，排斥的人多因“搞不清中国历史与人物”；从双语表现力的差异上分析，中文的张氏人物和场景堪称细腻，然而直译过去的英文则稍嫌琐碎；从人物关系上看，中文读者对复杂家族成员关系容易认同，而西方读者则易厌烦。（宋以朗、符立中 2013：53，71，84）

综上，从翻译研究的视阈审视张氏英文创作及出版结果，首先，我们不应怀疑张氏写出英美接受的英文作品的能力；其次，张氏对再现中国故事的坚持展现了东方作家的创作勇气，即便是屡遭拒稿、被迫不断改写，也不放弃注定被批判为“中国化”的叙事手法及被诟病为“语言缺陷”的成语俗语直译做法，堪称“先锋试验”。（同上：73）

5. 结论

《怨女》自译的特殊之处在于其英文版不唯一，且在英语语境出版命运截然相反。倘若*Pink Tears*没有沦为沧海遗珠，踪迹可寻、甚至是原稿通过某种途径、以某种形式完整呈现在读者面前，那将成为最理想的自译研究案例。通过对比*Pink Tears*和*The Rouge of the North*，我们将有机会探索文本差异，并在文本特征及出版境遇之间建立某种关联，得出的结论将在一定程度上启迪现代文学自译活动，因为使用英语创作的中国故事在英语语境得以出版，也是“中国故事走出去”文化战略的最终落实环节。*Pink Tears*虽然出版遭拒，但这种冷遇依然具有意义。虽然现阶段的析因只能借助作家本人的书信和言论推测和部分还原，但是它至少敦促翻译研究者推测遭拒的原因，同时认识到文学自译应充分考量出版的时代特点、政治因素、文化环境等，寻求在迎合英美读者群并寻求出版机会与保留中国故事

样貌的夹缝中间，如何调整文本、达成妥协，最终跨文化“走出去”。

自译双语文本创作时间和出版时间的区辨有助于确认语际方向，是自译文本特征研究的前提。语料库技术在通过抓取文本特征进而反思译本出版境遇方面的意义也不容小觑，借助该技术研究者能够从微观层面解读自译者在面对坚守创作初衷和迎合译入语读者口味两项任务时需要做出怎样的取舍决策；亦能够在宏观层面把握英美非翻译小说和我国自译者创作的英语小说之间的主题特色、语言偏好之间的异同，从而帮助推定译入语语境下出版市场需求，为文学自译者提供语言技术参考。在自译活动的委托人、出版方、读者、政治环境等不可控文本外因素恒定的情况下，自译者需要思考如何确保自己创作的英文文本能够被英美读者所接受，抓住两个语言世界，以免遗珠之恨。

注　释

1. 张爱玲文学遗产执行人宋以朗在书面文献中提及的该作品名均为*Pink Tears*，而非*The Pink Tears*。
2. 为确保对比的可行性，参照语料库选择20世纪40到70年代盛行的英美各3部小说：3部英国小说分别是《傲慢与偏见》、《简 · 爱》和《呼啸山庄》；3部美国小说为《乱世佳人》、《嘉莉妹妹》和《红字》。
3. 据陈亚明（2011），张爱玲在台北皇冠出版社1988年2月《续集》自序中否认《五四遗事》中英文版之间的翻译关系，因两版本均“迁就读者的口味”，故表现手法“略有出入”。
4. 据宋以朗、符立中（2013：169），麦卡锡（Richard McCarthy）系美国新闻处（USIS）文化部主任，1952年美新处彼时征选海明威《老人与海》的中文译者，张应征中选，遂与麦卡锡结识。张氏拿创作中的*The Rice-Sprout Song*前两章给麦卡锡，麦卡锡不但心仪且给当时赴港的马昆德看。张氏恰是马的中国书迷，张所著《十八春》（《半生缘》的前身）中部分情节借自马的小说*H. M. Pulman Esquire*。马返美后经自己的文学经纪人Marie Rodell（玛莉 · 勒德尔）女士推荐至相熟的Charles Scribner’s Sons出版。1955年该书出版，《纽约时报》等都给予好评，翻译版权卖出23种。
5. 据宋淇回忆，自五十年代马昆德来港结识张爱玲后，曾受张之托设法帮其“卖掉”包括*Spy Ring*在内的两篇短篇小说。马昆德也预测小说万一卖不掉，许是因为“读者不熟悉上海的背景”，并建议一旦遭拒可改投《纽约客》（*The New Yorker*）、《哈珀斯》（*Harper’s*）、《大西洋月刊》（*The Atlantic*）等刊，甚至可依照既定思路续写“约八个故事”成集，但坦言美国读者并不热衷“集结成书”的短篇小说。（张爱玲等 2011：52）
6. 宋淇、邝文美夫妇、麦卡锡、夏济安、夏志清等成为张爱玲海外发表的贵人。宋淇是1950年由麦卡锡聘于美新处译书部，1952年自张爱玲中选译者开始熟识。夏济安和宋淇是上海光华大学同窗、夏志清乃夏济安之弟。夏济安经宋淇赠阅《传奇》《留言》得识张爱玲（宋以朗、符立中 2013：168-169/172）。1966年中文版《怨女》在香港《星岛日报》连载是经宋淇介绍，出书事宜也曾在同年委托夏志清在访台时处理。

参考文献

Hsia, C. & J. Lau (eds.). 1971. *Twentieth-century Chinese Stories* [M]. New York: Columbia

University Press.

陈吉荣，2009，《基于自译语料的翻译理论研究——以张爱玲自译为个案》[M]。北京：中国社会科学出版社。

陈亚明，2011，张爱玲译事年表[J]，《新文学史料》(1)：64-69。

冯庆华，2008，《母语文化下的译者风格——〈红楼梦〉霍克斯与闵福德译本研究》[M]。上海：上海外语教育出版社。

刘绍铭，2013，《爱玲小馆》[M]。北京：海豚出版社。

宋以朗，2015，《宋家客厅：从钱钟书到张爱玲》[M]。广州：花城出版社。

宋以朗、符立中，2013，《张爱玲的文学世界》[M]。北京：新星出版社。

王德威，2004，《落地的麦子不死：张爱玲与“张派”传人》[M]。济南：山东画报出版社。

张爱玲、宋淇、邝文美，宋以朗编，2011，《张爱玲私语录》[C]。北京：北京十月文艺出版社。

周芬伶，2003，《艳异：张爱玲与中国文学》[M]。北京：中国华侨出版社。

通讯地址：030619 山西省晋中市太原师范学院外语系

汉译英新闻中间接引语时态不一致研究

浙江大学 郁伟伟

提要：鉴于新闻语篇时态不一致现象在信息传播中的重要性，本研究以汉译英翻译体新闻语篇中过去时报道动词与意图绝对现在时搭配现象为研究对象，对其时态特征和语义特征分布情况及导致时态不一致现象的句法、语义、语用层的翻译动因进行系统描写和分析。研究发现，过去时报道动词与意图绝对现在时搭配在数量、类型及语义特征上显著多于CNN本族语英语，存在过度使用的情况，而这是由于汉语原文在句法层、语义层和语用层“渗透”导致的，即“源语渗透效应”。

关键词：汉译英翻译体新闻语篇、时态不一致现象、分布特征、语义特征、翻译动因

1. 汉语时体划分

时态作为表示时间的语法形式，在英语、德语、法语等屈折变化语言中十分常见，但对汉语“有无时态”这一问题的讨论自20世纪50年代至今未间断过，大体分为三派：“有时有体”派认为汉语既有时范畴，也有体范畴，认为“时”主要通过时间副词、虚词等表示；“无时有体”派认为汉语没有表时间的形态标记和范畴，从语法上没有过去、现在和将来，但存在体范畴，“着”“了”等为表“体”的虚词，与“时”无关；“混合说”派承认汉语中时范畴的存在，将“体”标记和某些动词零形式视为“时”的形式标记。

2. 国内外间接引语时态不一致现象的相关研究

国内外很多研究者（Leech & Short 1981：326-327；辛斌2011）已注意到新闻语篇时态使用的不一致现象，有些学者称这种现象为“间接引语指示中心分离”。Declerck & Tanaka（1996）引入相对时态和绝对时态的概念来探讨这种现象。马景秀（2008）将此现象称为“新闻语篇间接引语时态非连续现象”。Vandelanotte（2006）、Davidse & Vandelanotte（2011）引入以报道说话人的时间零点为参照的第二个指称中心来分析此种语言现象。赖彦（2014，2015）称其为违背“逆移”规则的时态变异。虽然上述研究取得了不小成果，但缺少从语义趋向角度对时态不一致搭配现象的研究。胡开宝等（2018：27）也曾指出，语料库翻译语言研究

的瓶颈在于重语法、轻语义及语用。有鉴于此，本研究试图一方面扩展时态不一致搭配的语义趋向研究，另一方面拓展语料库翻译研究在语义、语用研究上新的可能性。

对间接引语时态不一致现象的解释原则主要包括Comrie（1986）的“持续有效性”原则、Declerck & Tanaka（1996）的真值条件、Smith（2007）的“双路径”从句时间独立及Davidse & Vandelanotte（2011）的引入被报道者的时间零点为参照的第二个指称中心。Comrie（1986：279，286）提出，如果间接引语内容上具有持续有效性，“逆移”原则是选择性的。持续有效性原则涉及本文讨论的语言现象，但由于并未进一步区分从句内不同时态，如现在完成时、现在时、将来条件时，而是都放在持续有效性下，因此本研究不考虑这一原则。

Declerck & Tanaka（1996）认为，相对时态是间接引语中非标记用法，而绝对时态是标记用法。他们声称使用相对还是绝对时态的标准是：如果报道从句中包含的命题对于被报道者来说是真实的，则使用相对时态；如果在“言语时刻是真实的，即在言语时刻世界还是真实的”，则使用绝对时态（转引自Davidse & Vandelanotte 2011：246）。Vandelanotte（2004）主要区分了直接、间接引语或思想表征忠于原报道的程度。Vandelanotte（2006）声称只有单数指称复杂性原型间接引语或思想表征及疏远型的间接引语或思想表征在报道者认识立场、态度或视角编码意义上屈从于主观化。Vandelanotte & Davidse（2009）认为应当废除传统的间接引语包含完全的指称转换及表达转换的观点，直接间接引语分析中应引入以被报道者的时间零点为参照的第二个指称中心。他使用Rigter（1982）的意图域这一概念，指“包含自身前提和真值条件的解释域，并且可参照前提和真值来评估和解释命题”。Davidse & Vandelanotte（2011）主要提出三点：一是报道动词在被报道者的话语中创造了第二个、替代的时间零点；二是提出了“意图绝对时”概念与被报道者的话语直接相关；最后，建议区分意图绝对时和真正绝对时，后者为语义事实并与当前报道者说话时间直接相关。

国内学者也对这一现象进行了探讨。马景秀（2008）运用Kiparsky（1968）的时间参数理论讨论了新闻语篇间接引语时态非连续现象，得出时态从属和独立时态共同作用造成了时态非连续现象。赖彦（2014：123-124）讨论了新闻语篇中过去时动词后跟现在时、现在完成时及现在将来时三种违背“逆移”规则的时态变异序列情况，总结了三种类型的特征以及语境变异条件，即“一般现在时”满足“命题内容的真值条件”“转述视角的转移”，“现在完成时”满足“当前的关联性”“事件的新近性”，“一般将来时”满足“将来时态的弱化与认知情态倾向的有机结合”。

总之，可以看出国内外对报道动词加现在时从句动词现象解释的相关研究尚

不充分，经历了从Comrie（1986）的“逆移原则”到Declerck & Tanaka（1996）的真值条件原则的转换。本文主要基于Declerck & Tanaka（1996）的真值条件原则，以Vandelanotte & Davidse（2009）、Davidse & Vandelanotte（2011）引入报道说话人时间零点对时态的划分标准为基础，对汉译英新闻语篇中过去时报道动词引导的间接引语时态不一致现象进行研究。

3. 语料来源和研究方法

国内时体翻译研究中鲜见针对间接引语时态选择问题的系统研究。因此，本研究试图从汉译英新闻话语间接引语时态使用入手，探究过去时报道动词与意图绝对现在时搭配使用的特征，并与英语本族语新闻语篇对比，从“源语渗透效应”角度探讨造成此类差异的原因。所用语料来自平行语料库及可比语料库，分别为：美国语言资源协会LDC出版的GALE汉英平行语料库（其中汉语新闻包括CCTV4中文国际频道和凤凰卫视新闻，约90万字，对应的英语译文语料约62万词），以及自建的CNN英语口语新闻语料库（约62万词），语料构成见表1、表2。

表1　GALE汉英平行语料库新闻及CNN新闻构成

语料库	语料来源	节目名称	时间跨度
GALE	CCTV4	CCTV4 Daily news	2004.06-2006.11
	CCTV4	CCTV4 news	2005.04-2006.01
	Phoenix	GLOBAL REPORT	2005.04-2005.12
	Phoenix	GOOD MORNING CHINA	2005.09-2006.01
CNN	CNN	Situation Room	2006.12-2007.12

其中，GALE汉英平行语料库分为三部分。第一部分包含38个文本，第二部分包含38个文本，第三部分包含34个文本。平行语料库三部分共计包含907,654汉字和628,053英文单词。第一部分链接为https://catalog.ldc.upenn.edu/LDC2007T23，第二、三部分汉语源文本的组成如下：GALE汉英平行语料库第二部分（引自LDC网站:http://www.ldc.upenn.edu/Catalog/catalogEntry.jsp?catalogId=LDC2008T08）

表2　GALE汉英平行语料库构成

名称	第一部分	第二部分	第三部分	总计
汉语原文	351,594	328,760	227,300	907,654
英文译文	240,481	227,532	160,040	628,053

在分析时态特征时，首先将汉译英英语译文语料库根据Vandelanotte（2005）标注方案扩展标注后进行多角度分析。Vandelanotte（2005）从COBUILD语料库中随机抽取了500例含有“said that”的句子，将其分为四类，分别是：relative tense、absolute past tense、absolute present tense: Speaker’s world as t0-world（即报道时间）、absolute present tense: non-Speaker’s world as t0-world。因此本研究基于上述分类方案，标注间接引语中的过去时报道动词及意图绝对现在时从句动词。

然后，与CNN英文母语新闻进行对比，分析英语译文中间接引语时态使用的不同特征，并在分析汉英平行语料库的基础上探究译文语料中时态不同特征是否是由源语的渗透效应导致的。

4. 过去时报道动词与意图绝对现在时的特征及原因分析

汉译英翻译体英语过去时报道动词与意图绝对现在时搭配共有1,458例，493种类型，本族语英语过去时报道动词与意图绝对现在时搭配共有278例，125种类型。二者在数量（p=0.000<0.05）和类型（p=0.000<0.05）上都存在显著差异。

通过过去时报道动词与意图绝对现在时从句动词的语义搭配类型，我们可以得到二者的语义分布频数。我们发现，翻译体英语和本族语英语过去时报道动词与意图绝对现在时搭配中，过去时报道动词有56种共有的语义类型，翻译体英语特有的语义类型有81种，本族语英语特有的语义类型有8种。对于意图绝对现在时从句动词，翻译体英语和本族语英语中共有的语义类型有58种，翻译体英语特有的语义类型有142种，本族语英语特有的语义类型有16种。

通过对比过去时报道动词引导的各个时态从句动词的语义类型，我们发现，译成绝对现在时动词的特有语义频数在2次及以上的有：[位置和方向]（4次）；[尊重]（4次）；[陆地上的车辆和运输]（3次）；[生命与生物]（3次）；[体育和游戏总称]（3次）；[衣服和个人物品]（3次）；[比较：种类]（2次）；[担心，忧虑，自信]（2次）；[话语行为]/[人]（2次）；[家具和家居]（2次）；[建筑物局部]/[船务、游泳等]（2次）；[商务：销售]（2次）。

接下来，我们将从句法层、语义层和语用层分别探讨时态不一致现象背后的

翻译动因。句法驱动分为五个方面：第一，视点体：了、着、过。第二，介词结构：在、于、自、以、把。第三，条件句：若、如果、假如、只要、万一、一、一旦、有无。第四，连词：否则、到、或者、即使、就算、虽然。第五，表比较：比、相比、较、比较、比起、变得、不如、更、相应。

黄敏（2006：50）指出，汉语缺乏专门的“时”形式，很多情况下要借助于“体”区别来表达“时”区别。经统计，过去时报道动词与意图绝对现在时的搭配的1,458例中，属于句法层驱动的共有89例。其中助词“了”出现38次，总结规律及例证可发现“了”在表达“被报道者或报道者认为是事实，并持承认和肯定态度”、“表示已发生事件产生的影响、具有的意义”和“表示对未来的还没有实现的假设、条件或对未来所持的态度”三个意义时常译作意图绝对现在时。“着”驱动的意图绝对现在时出现了28次，翻译动因总结如下，首先，“着”跟在实义动词之后，如“意味”“面临”等，表示动作的进行。其次，“着”跟在“期待”“盼望”等表将来可能性的能愿动词后，表示这种状态的持续。第三，“着”跟在“有”“存在”等表存在、带有的动词后，表示状态的持续。

条件句因为受真实条件句句型的管辖，所以使用了意图绝对现在时。连词“万一”“否则”表示对未来的假设，用意图绝对现在时表示将来。介词“在”主要有三种情况，一是“在”指具体的位于某个地理位置，是本义。其次，介词“在”表示抽象的“在于”，指的是办法或原因的追溯。第三，“在”表示时间上的“位于”，多用于“在……过程中”“在……情况下”。介词“以”表示用、采取，一般在接未实现的将来的一种意向时，译作意图绝对现在时。

语义驱动分为十个方面。第一，时间副词。第二，介词短语：在、于、自。第三，时间名词。第四，连系动词。第五，能愿动词。第六，终结性事件动词组合。第七，结果动词补语。第八，状态动词组合。第九，语义上叙实性词汇。第十，表“将来”的词汇。

语义层驱动有1,177例，其中能愿动词出现424次，如：要、可能、可以、能、能够、需要、必须、愿意、愿、无法、可能会等。经总结，能愿动词译作意图绝对现在时包括：can、may、should、must、have to、want、need、require、be able to、be capable of、enable、prove、intends、be incapable of等。

连系动词出现387例，其中“是”有271例，257例“是”及11例“为”译作is/are，11例“为”译作last。结合刘月华等（2001）对“是”字句的归类总结规律如下：首先，“是……的”表判断的句型有99例。112例译作is/are的“是”全为刘月华等（2001）中“是……的”句（二），“是”和“的”都是语气词，“多用来表示说话人对主语的评议、叙述或描写，全局往往带有一种说明情况、阐述道理、想使听话人接受或信服的肯定语气”（同上：771）。译作其他动词的14例“是”中有7例为“是……的”句（二），同样是表达对主语的评议、叙述或描写。这些

“是”或者不翻译，或者译作pose、signals、involve、dedicate、has等。其次，“是”表示前后等同或归类。第三，宾语对主语的某些方面进行说明。第四，用于说明、解释原因等，有时有申辩的意味。第五，表示肯定。

时间词中的时间副词有147例，其中“将”有15例译作意图绝对现在时。有些译作情态动词can、should，表示希望将来能实现的语气，有用一般现在进行时表将来的，有借助现在时动词表将来的，还有的译作一般现在时动词。“已经”有14例译作意图绝对现在时，其中译作is/are的表示被报道者认为是事实，另外，有借助现在时状态动词be prepared/ready表示状态一直持续的。“已经”“已”表示被报道者认为“已”后跟的为事实。“正在、正、在、依然、不断”译作现在进行时，表示动作正在进行或事件正在发生，译作be in contact with等介词短语，表示处于某种状态。“还”有4例，后跟表状态的或能愿动词，如“有”“存在”“年轻”“很多”“需要”“无法”译作意图现在时，表示状态还要持续。“尚”含否定意义的句子一般译作意图绝对现在时，后常跟否定副词“未”“不”“没（有）”“早”等。“一直以来”译作意图绝对现在时，表示动作或状态的持续进行。用于强调事实，或肯定语气时用意图现在时。“尽快”使用意图绝对现在时表达被报道者的语气。“暂时”表示其后的状态持续时间仅在被报道者的当下时刻，多含有否定意义。“后”表示只有“后”前的条件具备了，其后的动作和事件才能发生。“事先”表示一种条件。“先让”借助情态动词should表示事件的先后顺序。“进一步”译作意图绝对时，表示程度上肯定。“必将”表示肯定语气。“日益”表示时间上越来越接近。“再”表示程度上进一步。

时间词中的介词短语共有18例。这些译作意图现在时的介词短语，有时是被当作事实，有的是表示从被报道者的言语时刻开始。如果是表示将来时间，会借助情态动词或现在进行时等实现。时间词中的时间名词有71例。

终结性事件动词组合有52例，其中表示被报道者言语行为、立场态度的有：“赞成”“同意”“表示”“主张”。其他陈述客观事实的有：“符合”“接近”。语义上叙实性词汇有36例，多表示一些客观事实。如：“显示”“反映”“证明”“来自”。状态动词组合有37例，均表示主观感受，如：“认为”“感到”“喜欢”。结果动词补语有6例，如：“取得”“构成”“达到”。表“将来”的动词有4例，如：“企图”“提供”。副词有4例，“还”“只”强调程度或数量。

语用驱动分为两种，表示时态不一致时带有报道者特定意图。第一，语境。包括以下3种：数量词、指称语、其他。定义：只能通过上下文得出时态。第二，语气。包括以下6种：积极/消极动词、积极消极形容词、积极/否定副词、虚拟语气、“是……的”句、能愿动词的过去时。

语用层驱动有493例，其中与形容词“相关”的有2例。由虚拟语气引入的报道动词多为“要求”“建议”“推荐”“敦促”“强调”“提议”和“质疑”。终结

性事件动词“返回”和“忘”各1例。数量词中表惯常行为、频率的词有：“经常”“每天”“每日”“每次”“常”“普遍”和“第一次”。积极形容词有45例，如：“重要”“好”“高”“高兴”“满意”“密切”“巨大”和“良好”。消极形容词有14例，如：“低”“很难”。积极副词有9例，如：“都”“远远”和“正面”。否定副词有69例，如：“不”“没有”“未”“没”“不必”“别”和“无须”。积极动词有141例，“希望”“支持”“相信”“欢迎”“重视”“有利于”“坚持”“有望”和“期待”。消极动词有23例，如：“反对”“并非”。指称词有6例，如：“该”“这次”“这一次”“这样”“这个”。其他语用层次的词汇有112例，如“有权”“使”“有关”“视”和“无权”等。

5. 结语

研究以汉译英翻译体新闻语篇中间接引语的时态不一致现象为研究对象，对间接引语中过去时报道动词与意图绝对现在时从句动词不一致搭配的时态特征和语义特征分布情况，及导致这种时态不一致现象的句法、语义、语用层的翻译动因进行了系统的描写和分析。

以上基于语料库的实证分析有以下主要发现：

（1）汉译英翻译体英语过去时报道动词与意图绝对现在时搭配种类（493类)和本族语种类（125类）差异显著，译文（1,458例）与本族语（278例）数量存在显著差异。（2）汉译英翻译体英语和本族语者英语过去时报道动词有56种共同的语义，译文过去时报道动词特有的语义类型有81种，本族语去时报道动词共有8种特有的语义类型；（3）汉译英翻译体英语和本族语者英语意图现在时从句动词有58种共同的语义，译文英语意图现在时从句动词特有的语义类型有142种，本族语英语意图现在时从句动词共有8种特有的语义类型。

“源语渗透效应”是造成汉译英新闻中过去时报道动词与意图绝对现在时从句动词时态不一致过度使用的背后原因。其他时态不一致的搭配类型，如：过去时报道动词与意图绝对过去时搭配、过去时报道动词与意图现在完成时搭配、过去时报道动词与意图将来时搭配、过去时报道动词与真正绝对现在时搭配、过去时报道动词与真正现在完成时搭配、过去时报道动词与真正现在将来时搭配等的类型、语义特征及翻译动因还有待进一步考察。

参考文献

Comrie, B. 1986. Tense in indirect speech [J]. *Folia Linguistica* 20(3-4): 265-296.

Davidse, K. & L. Vandelanotte. 2011. Tense use in direct and indirect speech in English [J]. *Journal of Pragmatics* 43(1): 236-250.

Declerck, R. & K. Tanaka. 1996. Constraints on tense choice in reported speech [J]. *Studia Linguistica* 50(3): 283-301.

Kiparsky, P. 1968. Linguistic universals and linguistic change [A]. In B. Emmon & R. Harms (eds.). *Universals in Linguistic Theory* [C]. New York: Holt, Rinehart, and Winston. 170-202.

Leech, G. & M. Short. 1981. *Style in Fiction: A Linguistic Introduction to English Fictional Prose* [M]. London: Longman.

Rigter, B. 1982. Intentional domains and the use of tense, perfect and modals in English [J]. *Journal of Semantics* 1(2): 95-145.

Smith, C. 2007. Tense and temporal interpretation [J]. *Lingua* 117(2): 419-436.

Vandelanotte, L. 2004. Deixis and grounding in speech and thought representation [J]. *Journal of Pragmatics* 36(3): 489-520.

Vandelanotte, L. 2005. Tense in indirect speech or thought: Some proposed modifications [A]. In B. Hollebrandse, A. van Hout & C. Vet (eds.). *Crosslinguistic Views on Tense, Aspect and Modality* [C]. Amsterdam: Rodopi. 61-75.

Vandelanotte, L. 2006. Speech or thought representation and subjectification, or on the need to think twice [J]. *Belgian Journal of Linguistics* 20: 137-168.

Vandelanotte, L. & K. Davidse. 2009. The emergence and structure of be like and related quotatives: A constructional account [J]. *Cognitive Linguistics* 20(4): 777-807.

胡开宝、朱一凡、李晓倩，2018，《语料库翻译学》[M]。上海：上海交通大学出版社。

黄　敏，2006，《论现代汉语语篇中“时”的表达及功能》[M]。南昌：江西人民出版社。

赖　彦，2014，新闻报道语篇的时态序列变异及认知阐释 [J]，《浙江传媒学院学报》(6)：121-126。

赖　彦，2015，新闻语篇间接转述言语的时体变异 [J]，《外语与外语教学》(1)：19-25。

刘月华、潘文娱、故　韡，2001，《实用现代汉语语法》[M]。北京：商务印书馆。

马景秀，2008，英语新闻语篇间接引语时态非连续性现象刍议 [J]，《西安外国语大学学报》(4)：18-21。

辛　斌，2011，间接引语指示中心的统一和分离：认知符号学的视角 [J]，《外语研究》(3)：7-11。

通讯地址：310058 浙江省杭州市 浙江大学外国语言文化与国际交流学院

学习者学术英语写作中作者身份凸显度研究*

河南师范大学 王 莉 娄宝翠

提要：学术语篇中作者身份有助于实现语篇交际功能，而引述句是凸显作者身份的重要语言资源。本研究通过与国际期刊论文语料对比，聚焦引述句主语，探究学习者在学术英语写作中作者身份凸显度特征以及在本科、硕士和博士不同阶段的发展性特征。研究结果表明，学习者总体上身份凸显意识薄弱。从发展特点上看，这种学术身份凸显随学习者的学术水平提高，总体呈上升趋势，但在具体使用上存在很大差异：本科生存在使用不规范现象，身份凸显意识薄弱；硕士生倾向于借助他人研究来增强语篇可信度，有利于凸显作者身份；博士生总体上呈现了相对成熟的作者身份。

关键词：作者身份、引述句、主语、学术英语写作

1. 引言

学术写作表达观点具有客观性和科学性，应避免出现作者对语篇的主观介入（Gong & Dragga 1995；Hyland 2001；Lester 1993；Spencer & Arbon 1996）。然而最近研究表明，作者身份（authorial identity）是学术写作中不可缺少的成分，学术写作是作者身份构建的重要场所，二者之间存在很强的互动关系（Çandarlı *et al.* 2015；Rahimiv & Kuhi 2014），因为学术写作的作者在展现自己研究成果客观性的同时，也会表明自己的立场、构建自己的身份以获得读者认同。国内外学者有关作者身份的研究大多数集中在自我提及语（self-mentions）这种极为显性的手段上（如Hyland 2002；Ivanič 1998；Kuo 1999；Tang & John 1999；高霞 2015；吴格奇 2013；王晶晶、吕中舌 2017）。但除了自我提及语，还可通过隐性的手段或无生命的抽象主体来凸显作者身份。因此，本文聚焦引述句主语，分析学习者在学术写作中作者身份凸显特征以及在本科、硕士、博士等不同阶段的发展性特征，以期为学术英语教学和写作带来一定启示。

* 本研究系国家社科基金项目“基于语料库的学术英语元话语特征对比研究”（14BYY150），2015年度河南省科技创新人才支持项目，2017年研究生创新项目（YW201701）“学习者学术写作中的作者身份构建研究”，“学习者学术英语与作词块使用模式研究”阶段成果。

2. 文献综述

学术语篇是作者参与的社会化行为（Hyland 2002），用于凸显作者身份的语言手段有多种，如模糊语和引述句等，其中引述句作为确立作者身份的重要语言资源被大多数学者忽视。恰当地使用引述句有利于作者在学术语篇中表达立场，呈现作者身份。以往有较多研究关注引述动词（见Manan & Noor 2014；Hyland & Tse 2005；Hunston 1993；Thompson & Ye 1991；Yeganeh & Boghayeri 2015）。其中Thompson & Ye（1991）把引述动词分为语篇动词（textual verbs）、心理动词（mental verbs）和研究动词（research verbs）三类。Hyland & Tse（2005）基于跨学科角度对期刊论文摘要中的评价型that结构进行分析，结果表明，硬学科中与之高频搭配的引述动词是研究动词（如show、demonstrate），软学科中的高频搭配动词是心理动词（如suggest、argue）。

小句的主语是表达人际意义语气结构的重要组成部分（Halliday 1994）。因此引述句主语的恰当使用，能够帮助作者在语篇中呈现自己身份，实现学术语篇中的交际功能（鞠玉梅 2016）。Charles（2006）对比研究了自然学科和社会学科中的引述句特征，根据来源将其分为自引句和他引句，在此基础上，根据主语分布规律又将句子主语分为人（Human）、非人（Non-human）和it结构。武姜生（2010）通过分析中国英语专业和美国本科生毕业论文语料库，发现学习者对客观性引述句主语使用不足，过度使用主观性引述主语，身份凸显意识较高。鞠玉梅（2016）分析了英语本族语者、中国学习者和英语专业硕士生中引述句主语特征，研究发现中国学习者身份凸显意识薄弱，硕士生虽具有身份构建意识但还与本族语者存在较大差距。不过这些研究都单一地从静态视角出发，本文以国际期刊论文为参考语料，从静态和动态视角考察学习者利用引述句主语凸显作者身份的特征。

本研究参考Charles（2006）的部分研究内容，此外，名物化形式（见武姜生 2010）也在考查范围之内，例如在句子the theory operates under assumption that a shared reality is psychologically essential for human existence中，assumption与后面的that从句构成名物化引述句，因此本研究理论框架建立在Charles（2006）和武姜生（2010）基础上，把it结构和名物化形式归于“非人”主语（见表1）。

表1　学习者身份显现度理论框架

引述句主语	自我引述句 self-sourced reports	他人引述句 other-sourced reports
人 Human	第一人称代词（如we、I）	第三人称代词、人名（如Hyland、Chomsky、she、he、they）
	自我指代名词（如the author、writer、researcher）	
非人 Non-human	研究名词（如the study、data、result、findings）	泛指他人研究名词（如the previous studies、some studies）
	语篇名词（如section 1、chapter 2、the thesis、paper）	it结构（如it is well known that…）
	it结构（it is suggested / certain that…）	名物化形式（如the model works under assumption that…）

3. 研究设计

3.1　语料来源

本文采用课题组自建的两个语料库，包括英语专业学习者学位论文语料库和国际学者发表的国际期刊论文语料库，二者均为应用语言学方向的论文。其中英语专业学习者论文语料库包括：来自于5所高校的72篇本科生学位论文语料库（简称BA_C），库容399,840；13所高校硕士生2011—2013年的25篇学位论文语料库（简称MA_C），库容384,176；以及5所高校博士生2008—2013年的10篇学位论文语料库（简称PHD_C），库容499,866。国际期刊论文语料库（简称IJA_C）来源于*Journal of Pragmatics*、*English for Specific Purposes*、*Journal of English for Academic Purposes*、*Journal of Second Language Writing*、*Language & Communication*、*Linguistics and Education*、*Applied Linguistics*、*TESOL Quarterly*等8种国际期刊论文50篇，库容370,516。上述语料库文本数据都只保留标题、摘要和正文信息。

3.2　研究问题

本研究以国际期刊论文为参照，考察学习者英语学位论文中引述句主语分布特征和发展性特征，探究学习者如何凸显作者身份以及这种显现度呈现何种趋势。具体的研究问题为：

（1）与国际学者相比，学习者引述句主语分布有何特征，对身份呈现度有何影响？

（2）学习者在本科、硕士和博士阶段呈现的身份特征有何变化？

3.3　研究步骤

首先，利用WordSmith 6.0工具，参照Charles（2006）和Hyland & Tse（2005）的研究方法，即首先检索that索引行。虽然that可能存在省略情况，但前人研究（Biber *et al.* 1999）表明包含that的转述句是学术写作的规范。因此本研究仅关注未省略that的引述句主语特征。

其次，检索完成后逐条剔除不符合条件的语料，如that做指示代词和关系代词的情况。然后对已筛选的引述句进行人工分类，具体操作中先由作者本人进行判断，再由课题组其他成员复核，经讨论达成一致意见后确定最终分类。

最后，统计分析按标准频数和原始频数报告结果，标准频数采用的是每百万词的统计方法，并运用梁茂成开发的对数似然率计算工具（Loglikelihood，简称LL）检验对比频数间的差异。

4. 研究结果与讨论

4.1　引述句主语分布特征

统计结果显示，引述句主语的分布总体上在学习者英语硕士论文和国际期刊论文中无明显差异（LL=1.554，p=0.213>0.05)（见表2），说明学习者总体上试图通过引述他人观点来增加语篇可信度，有利于培养学术自信，构建学术身份。

表2　学习者引述句主语分布特征

引述句主语			学习者语料库	国际学者语料库	LL值	p值
			标准频数	标准频数		
自我引述	人	第一人称代词	393.34	439.93	-1.514	0.218
		自我指代名词	53.74	5.40	22.753	0.000
		合计	447.08	445.33	0.002	0.964
	非人	研究名词	391.00	658.54	-41.677	0.000
		语篇名词	9.35	5.40	0.587	0.444
		it 结构	968.16	593.77	48.319	0.000
		合计	1359.16	1252.31	2.668	0.102

（待续）

（续表）

引述句主语			学习者语料库	国际学者语料库	LL值	*p*值
			标准频数	标准频数		
他人引述	人	第三人称代词、人名	906.63	809.68	3.118	0.077
	非人	泛指他人研究名词	66.98	126.85	-11.490	0.001
		it结构	24.92	8.10	4.729	0.030
		名物化形式	123.06	164.64	-3.562	0.059
		合计	214.96	299.59	4.729	0.030
总计			2927.83	2806.91	1.554	0.213

4.1.1　自我引述句主语特征

表2显示，以“人”做主语的频数远低于以“非人”做主语的频数，体现了学术写作的客观性和科学性，为呈现作者身份创造了条件。“人”做主语时，学习者与国际学者无明显差异（LL=0.002，p=0.964>0.05），说明学习者作为学术写作新手，有意模仿国际学者写作风格。主语是命题责任（propositional responsibility）的承担者（Groom 2000），第一人称代词做主语是凸显作者身份、承担命题责任最直接和最显著的方式（Hyland 2001；Kuo 1999；Tang & John 1999）。进一步观察索引行发现，国际学者较多使用第一人称代词单数形式I，凸显了作者作为研究者的核心地位，展现了研究者专业学术身份（如例1）；学习者绝大部分使用复数形式we，试图邀请读者参与到语篇之中，减少自己所需承担的责任（如例2）。这进一步说明学习者有意模仿国际学者写作规范，但作为初学者缺乏学术自信，不敢像国际学者那样直接凸显作者身份。值得注意的是，与国际学者相比，学习者过度使用自我指代名词（LL=22.753，p=0.000<0.05），频数约为国际学者的4倍，因为学习者通过大量使用 the author代替第一人称代词，表明学习者在学术写作中多以旁观者的身份出现，这一定程度上隐藏了作者身份（如例3）。

（1）**I** argue that texts allow writers to persuade readers and meet…

（2）In Chapter 3 **we** mention that generally there are two cultural…

（3）In the sentence, **the author** states that the designs of the dresses are worthy of his or her compliment…

以“非人”做主语时，与国际学者相比，学习者过度使用it结构（LL=48.319，p=0.000<0.05)，显著少用研究名词（LL=-41.677，p=0.000<0.05)。在学术写作中，使用it结构可以表达作者观点而不用提及作者本人，很大程度上降低了“作者能见度（writer’s visibility）”（Hyland 2002；Charles 2006）。例（4）中学习者通过使用缺乏明确引述来源的it做主语，隐藏了作者身份。研究名词表示研究过程、结果和发现之类的名词，以研究名词做主语时，研究结果（Results）和发现（Findings）被赋予命题责任的承担者，凸显了“研究事实说话”的写作特点，说明国际学者凸显作者身份的同时，也注重学术语篇的客观性，展现了其成熟的写作风格和学术权威（如例5），学习者显著少用名词主语说明他们还需提高学术规约意识。

（4）**It** is suggested that a better understanding of ways of conveying…

（5）**Results** suggest that these miscues do indeed significantly…

学习者总体上有意模仿国际学者的写作规范，主观介入学术语篇凸显作者身份，但在具体使用上存在很大差异；学习者学术水平有限，进行学术写作者时缺乏自信心，不能准确把握呈现作者身份和语篇客观性之间的关系。

4.1.2　他人引述句主语特征

以“人”做主语的频数远高于以“非人”做主语的频数，前者主要用于引述前人的观点，以引述者的姓名或人称代词的形式呈现，比后者的指代意义更为明确。说明学习者有意识将自己的研究置于学术领域，提高了语篇可读性和权威性。但与国际学者相比，学习者使用得更多，可能是因为他们试图通过大量地引用“权威人士”的观点，来增加自己观点的可信度，也可能是因为作为学习者，学术写作水平有限，想通过引用文献来凸显自己在话语社团内的一席之地（如例6、7）。

（6）**Leeuwen** (1999, 2006) assume that image, color, music, typography and other visual modes are similar to language and they fulfill the…

（7）**Wang Xijie** (1989) said that a pun typically reflects the comprehensiveness of rhetoric and is a common issue which draws the attention…

以“非人”做主语时，学习者过度使用“it 结构”（LL=4.729，p=0.030<0.05）过少使用“泛指他人研究名词”（LL=-11.490，p=0.001<0.05）。例（8）显示，学习者通过陈述大家共享的常识，并未涉及作者的身份，在某种程度上导致语篇缺乏权威性。“泛指他人研究名词”做主语时，在某种程度上凸显作者对话语社团研究情况的了解程度，有助于让读者接受作者的学术身份。例（9）通过展现前人的研究的成果，为自己的研究做铺垫，有利于增强语篇可信度。然而学习者存在使用不足现象，说明只有少数学习者掌握该用法。值得注意的是，“名物化形式”的引述句结构难度大于“it结构”，学习者相对少用该结构可能是和其语言水平有关。

（8）**It** is generally believed that Appraisal theory was firstly introduced into China by Prof. Wang Zhenhua…

（9）Many previous **studies** show that stance adverbial is one of the most common devices of conveying stance and stance.

学习者有意构建作者身份，但在具体的学术写作中不够成熟。他们过度强调语篇的权威性和客观性，虽然这有利于培养学术自信，展现学术身份，但学习者一直处于准备构建状态，有待进一步向国际学者靠拢。

4.2 引述句主语发展性特征

统计结果显示，在自我引述和他人引述句中，学习者在不同阶段使用特征存在较大差异（见表3）。

表3 学习者引述句主语发展性特征

引述句主语			本科生			硕士生			博士生		
			原始频数	标准频数	百分比	原始频数	标准频数	百分比	原始频数	标准频数	百分比
自我引述	人	第一人称代词	187	486.76	33%	184	460.18	25%	134	268.07	13%
		自我指代名词	15	39.04	2.7%	20	50.02	2.7%	34	68.02	3.3%
	非人	研究名词	90	234.27	16%	126	315.13	17%	286	572.15	28%
		语篇名词	7	18.22	1.2%	4	10.00	0.5%	1	2.00	0.2%
		it 结构	266	692.39	47%	417	1042.92	56%	560	1120.30	55%
	合计		565	1470.68	-	751	1878.25	-	1015	2030.54	-
他人引述	人	第三人称代词、人名	251	653.35	85%	394	985.39	88%	519	1038.28	74%
	非人	泛指他人研究名词	7	18.22	2.4%	14	35.01	3.1%	65	130.04	9.3%
		it结构	9	23.43	3.1%	7	17.51	1.6%	16	32.01	2.3%
		名物化形式	26	67.68	8.9%	34	85.03	7.6%	98	196.05	14%
	合计		293	762.68	-	449	1122.94	-	698	1396.38	-
总计			858	2233.36	-	1200	3001.19	-	1713	3426.91	-

4.2.1　自我引述句主语的发展性特征

表3显示，在自我引述句以“人”和“非人”做主语时，从百分比来看，学习者的总体趋势呈现出不规则变化趋势。值得注意的是以“人”做主语中第一人称代词做主语情况。本科生所占比例最高（33%），其次是硕士生（25%），博士生最低（13%）。说明总体上学习者随着学术水平提高，身份凸显度呈现下降趋势，但分析语料发现学习者在具体用法上身份凸显意识呈现相反趋势。学习者都倾向使用作者身份凸显度较低第一人称代词复数we，但搭配结构存在很大差异：本科生论文中过多用we know + that结构，硕士生和博士生都较多使用we find (found)+that结构，不同之处在于硕士生多用一般现在时结构，而博士生多用过去式结构，表示动作已经完成需要作者承担明确责任，如：

（10）**We** know that many people go out to earn a living, leaving…（BA_C）

（11）From what has been discussed above, **we** find that there is a…（MA_C）

（12）Furthermore, **we** found that the analytical tools in the…（PHD_C）

上述例句表明，学习者在不同学习阶段，对we不同搭配结构呈现出不同的语篇功能，Tang & John（1999）根据身份凸显度由弱到强将第一人称代词分为六种语篇功能：代表、向导、建筑师、叙述者、意见持有者和创始者。例（10）中we承担“代表”的语篇功能，即代表所有的读者，作者身份凸显度最低；例（11）中we承担“向导”的语篇功能，即通过向读者展现已有知识，引领读者通读全文，作者身份凸显度略高于“代表”；例（12）中we承担“创始者”的语篇功能，即作者呈现自己的新思想、新发现，作者身份凸显度最高。由此说明，学习者随学术水平提升，学术自信逐步增强，身份凸显度有所提升。

在“非人”做主语的情况中，学习者使用it结构呈现不规则趋势，从本科到硕士阶段明显提升，到博士阶段略微下降，研究名词的比例呈现上升趋势，而语篇名词的比例呈现下降趋势。值得注意的是使用频数最高的it结构，观察索引行发现，本科生多使用“It can be seen/said that…”，其表达形式单一，偏向口语化，文章缺乏权威性（如例13），硕士生和博士生多用“it is（has been) found / suggested / argued / assumed that...”结构，显得较为正式或礼貌，增强文章可信度，为构建作者身份创造了条件（如例14、15），学习者身份构建意识有很大提升。到博士阶段略微下降可能是因为博士生已突破身份构建的“准备状态”，开始使用更直观的凸显身份的表达手段（如人称代词）。

（13）…**it** can be seen that in both China and English-speaking countries…（BA_C）

（14）…**it** is found out that professional women take on a unique…（MA_C）

（15）…**it** has been suggested that some task types result in poor quality…（PHD_C）

学习者自我引述句主语的发展特征较为复杂，因此频数比例不是判断作者身

份凸显度的唯一方式。学习者学术水平越高，身份凸显意识越强，尤其博士阶段表现更为突出。

4.2.2　他人引述句主语的发展性特征

他人引述句中，“人”和“非人”做主语时，从百分比来看，学习者总体趋势呈现不规则现象。以“人”做主语时，包括第三人称代词和人名两种情况，相比之下，后者指代的意义更为明确。总体上，学习者从本科阶段（85%）到硕士（88%）阶段略微上升，到博士阶段（74%）呈现明显下降趋势。分析语料发现，本科生和硕士生较多使用“第三人称代词”做主语，引述来源不明确，降低了读者的可接受性，不利于劝说读者接受作者观点（如例16、17）；而博士生多用人名做主语，意义指代明确，展现了较为成熟的专业知识，容易让读者信服，与读者产生共鸣，从而呈现恰当的作者身份（如例18）。值得注意的是，人名做主语时，本科生存在使用不规范现象，他们引用“专家”的观点时，未标注年份，很大程度上降低语篇的可信度，这种现象在硕士论文中极少出现（如例19），在博士论文中没有出现。

（16）**He** maintains that anxious writers should be encouraged to…（BA_C）

（17）**They** also consider that Chinese government keeps emphasizing…（MA_C）

（18）**Kaplan** (1966: 14) posits that every language and culture have its…（PHD_C）

（19）**Tytler points out that** if a translated work cannot take on the…（BA_C）

以“非人”做主语时，名物化形式和it结构呈现先下降后上升趋势，泛指他人研究名词呈现上升趋势，值得一提的是使用频数最高的名物化形式，该表达增强了语篇的客观性。本科生较多使用名物化形式可能就是为了避免直接凸显作者身份而刻意模仿国际学者的一种写作策略。观察索引行发现，学习者使用类型较为单一，多以“…assumption / idea / understanding that…”形式出现（如例20）。硕士生阶段使用频数有所下降，可能是因为他们意识到凸显作者身份的重要性，有意减少这种明确性的客观表达，而到博士阶段这种结构使用频数明显上升，可能因为名物化形式句法结构难度大于it结构，随着他们语言水平提高，对这种较难结构掌握较为熟练，急于通过语篇展现自己的学术功底。此外，选词丰富度远大于本科生，如多以“…view / idea / claim / belief / assumption / argument that…”形式出现（如例21）。

（20）Goldberg projects a very important **assumption** that a difference…（BA_C）

（21）…Wodak mentioned the **argument** that SFL has not been…（PHD_C）

学习者在不同阶段的主语使用特征差异很大，本科生存在使用不规范现象，身份凸显意识薄弱；硕士生处于学术写作水平发展的过渡阶段，身份凸显意识提高；博士阶段学术水平提升，但过于刻意使用较为复杂的名物化句式结构，而忽视

作者身份的凸显。再次说明学习者未能熟练协调语篇客观性和主观性之间的关系。

5. 总结

本文主要考察了学习者引述句主语使用特征以及其在不同阶段的发展性特征，目的是探究学习者如何凸显自我身份及在不同阶段身份显现度呈现何种趋势。通过以上研究发现：与国际学者相比，学习者总体上身份凸显意识薄弱，不注重构建自己的作者身份，在自我引述句中有意避开身份凸显度过高的人称代词，寻求隐藏作者身份的表达手段。从发展的角度来看，随着学术水平提高，学习者总体上身份凸显意识增强，但在具体使用上存在很大差异：本科生对引述句主语的使用存在句式单一、使用不规范现象，较多使用缺乏明确引述来源的主语，自我观点表达和与话语社团内的互动意识薄弱；博士生则在语言使用形式上较为多样化和规范化，善于表达自己观点，实现与话语社团内成员的互动，展现了较为成熟的学者身份形象；硕士生处于本科生和博士生之间的过渡阶段，语言形式比本科生更正式，更趋于借助前人的研究来增强自己语篇的可信度，为呈现作者身份做准备。

本研究可为科研工作者和学术英语教师提供启示，促使其科研论文的写作及在教学中加强学生通过使用引述句主语凸显作者身份的意识。可以根据学习者所处的不同阶段，有针对性地培养其对引述句主语和作者身份凸显度之间联系的正确理解：（1）本科阶段注重培养学习者语言表达的规范，为凸显作者身份做准备；（2）硕士阶段注重培养其学术写作的连贯及水平，提高学习者勇于凸显作者身份的自信；（3）博士阶段有目的地培养学习者向国际学者靠拢，构建成熟的作者身份。本文也存在一些不足。呈现作者身份的手段有很多，本研究只对其中一种进行了研究，因此，后续有必要对其他呈现作者身份的手段进行更深入的研究。

参考文献

Biber, D., S. Johansson, G. Leech, S. Conrad & E. Finegan. 1999. *Longman Grammar of Spoken and Written English* [M]. London: Longman.

Çandarlı, D., Y. Bayyurt & L. Martı. 2015. Authorial presence in L1 and L2 novice academic writing: Cross-linguistic and cross-cultural perspectives [J]. *Journal of English for Academic Purposes* 20: 192-202.

Charles, M. 2006. The construction of stance in reporting clauses: A cross-disciplinary study of theses [J]. *Applied Linguistics* 27(3): 492-518.

Groom, N. 2000. Attribution and averral revisited: Three perspectives on manifest intertextuality in academic writing [A]. In P. Thompson (ed.). *Patterns and Perspectives: Insight into EAP Writing Practice* [C]. Reading: University of Reading. 15-25.

Gong, G. & S. Dragga. 1995. *A Writer's Repertoire* [M]. New York: Longman.

Kuo, C. 1999. The use of personal pronouns: Role relationships in scientific journal articles [J].

English for Specific Purposes 18(2): 121-138.

Manan, N. & N. Noor. 2014. Analysis of reporting verbs in master's theses [J]. *Procedia-Social and Behavioral Sciences* 134: 140-145.

Halliday, M. 1994. *An Introduction to Functional Grammar* [M]. London: Edward Arnold.

Hyland, K. 2001. Humble servants of the discipline? Self-mention in research articles [J]. *English for Specific Purposes* 20(3): 207-226.

Hyland, K. 2002. Authority and invisibility: Authorial identity in academic writing [J]. *Journal of Pragmatics* 34(8): 1091-1112.

Hyland, K & P. Tse. 2005. Hooking the reader: A corpus study of evaluative that in abstracts [J]. *English for Specific Purposes* 24(2): 123-139.

Hunston, S. 1993. Professional conflict: Disagreement in academic discourse [A]. In M. Baker & G. Francis (eds.). *Text and Technology: In Honour of John Sinclair* [C]. Amsterdam: John Benjamins. 115-134.

Ivanič, R. 1998. *Writing and Identity* [M]. Amsterdam: John Benjamins.

Lester, J. 1993. W*riting Research Papers (Seventh Edition)* [M]. Norwood: HarperCollins.

Rahimiv, M. & D. Kuhi. 2014. An exploration of discoursal construction of identity in academic writing [J]. *Procedia-Social and Behavioral Sciences* 98: 1492-1501.

Spencer, C. & B. Arbon. 1996. *Foundations of Writing: Developing Research and Academic Writing Skills* [M]. New York: McGraw-Hill.

Tang, R. & S. John. 1999. The "I" in identity: Exploring writer identity in student academic writing through the first person pronoun [J]. *English for Specific Purposes* 18 (s1): S23-S39.

Thompson, G. & Y. Ye. 1991. Evaluation in the reporting verbs used in academic papers [J]. *Applied Linguistics* 12(4): 365-382.

Yeganeh, M. & M. Boghayeri. 2015. The frequency and function of reporting verbs in research articles written by native Persian and English speakers [J]. *Procedia-Social and Behavioral Sciences* 192: 582-586.

高　霞，2015，基于中外科学家可比语料库的第一人称代词研究 [J],《外语教学》(2)：30-34。

鞠玉梅，2016，学术写作中引述句的主语特征与身份构建研究 [J],《外语教学与研究》(6)：926-936。

王晶晶、吕中舌，2017，理工科博士生学术英语写作中的作者自我指称语研究 [J],《外语界》(2)：89-96。

吴格奇，2013，学术论文作者自称与身份构建——一项基于语料库的英汉对比研究[J],《解放军外国语学院学报》(3)：6-11。

武姜生，2010，大学生英语学术写作中引述句的主语特征 [J],《中国外语》(2)：27-32。

通讯地址： 4530007 河南省新乡市河南师范大学外国语学院

BioDEAP生命科学学术英语语料库的创建

中国科学院大学 彭 工

提要：本文介绍BioDEAP生命科学学术英语语料库的总体设计思路以及具体实施步骤。对建库原则、语料采集、文件命名、文本转换及清洁、元信息及文本结构信息标注方式、软件工具操作方法等方面做了说明。文章最后探讨了生命科学学术英语料库的后期开发及应用。

关键词：生命科学、学术英语、语料库建设

1. 引言

多学科交叉是现代生命科学的重要特征。生命科学学术英语语料库建设对了解其语言表述规范、促进中国学术成果的国际发表、英语教学的语境化实证研究都具有重要的理论意义与应用价值。语料库按照所选取语料的特点可分为通用语料库、专用语料库、单语语料库、平行语料库及多语种语料库等不同类型。通用语料库一般较大，用来描写某一语言的全貌（梁茂成等 2010）。专用语料库通常较小，并不旨在全面代表某种语言的整体，而是只反映该语言的特定部分，这其中一个重要的专业研究领域是学术英语（Lee 2010）。

从20世纪80年代开始，随着计算机技术的发展、网络技术的广泛运用，语料库相关研究取得了令人瞩目的成就。“语料库开发正在向两头快速发展和延伸”：“一是通用型的、基于网络的超大型语料库开发”，“二是个性化、专门化、行业化的小型语料库开发”（桂诗春等 2010）。国外大型、超大型的通用语料库有英国国家语料库（BNC）、美国当代英语语料库（COCA）等。专门化语料库中，发展比较迅速而且影响力较大的是学术英语语料库。继美国密歇根大学的学术口语语料库（MICASE）之后，各国家地区同类语料库相继建成，包括英国学术英语口语语料库（BASE）、利默里克-贝尔法斯特学术英语口语语料库（LIBEL CASE）、香港城市大学学术英语口语语料库（CUCASE）、香港理工大学口语语料库（HKCSE）。随后，密歇根高水平学生论文库（MICUSP）、英国学术书面英语语料库（BAWE）、香港理工大学研究论文库2007（CRA 2007）、香港理工大学期刊文章语料库（CRA 2014）等一批书面语学术语料库相继建成。个人自建学术语料库有Coxhead的学术语料库（Coxhead’s Academic Corpus）、Hyland

的期刊论文和学术访谈语料库。在中国大陆地区，继上海交通大学科技英语语料库（JDEST）之后，许多机构或个人也相继建设了基于本校或个人教学、科研需求的专用学术英语语料库。例如广州石油大学的“石油英语语料库”、对外经济贸易大学的“商务英语语料库”、黑龙江大学的“商务英语语料库”、大连海事大学的“海事英语语料库”以及东华大学的“东华大学科技英语语料库”等。

语料库的出现对语言研究产生了巨大影响（何中清、彭宣维 2011）。国内学者于1992—2015年24年间在核心期刊上发表的语料库语言学相关研究论文数量达3,128篇，涵盖语言教学、语料库理论研究、语义研究、翻译研究、词汇研究、语法研究等多个领域（张新杰 2017）。然而，国内生命科学领域的语料库相关研究还不够深入。迄今为止，在中国知网检索到相关论文11篇，其中9篇与农业学科相关，6篇讨论语料库建设。如范晶晶、李丽霞（2014）提出了创建农业学术英语语料建设构想；吴蕾等（2014）探讨了如何建立科技语料库，举例说明了生命学科语料库的目录及文档命名方式；王敏、李丽霞（2014）以及李丽霞、白璐（2016）自建涉农新闻小型语料库，做了名词化、被动语态、主题词语义场等方面的语言特点分析。刘佳、韩丽娜（2014）基于自建语料库考察了Coxhead（2000）编制的通用学术词表在环境科学专业英语学习中的适用性；刘萍等（2015）探讨了农科学术英语论文语料库的创建及其在博士生和本科生学术写作教学的应用；陶玲（2017）指出建立小型专业化语料库对农业科技期刊英文摘要编写十分重要。

以上分析表明，生命科学学术英语语料库存在重复建设、复用率低的问题。现有语料库资源，特别是某些院校建设的较大规模的科技语料库子库，由于知识产权等各种原因，不对外开放，难以实现资源共享。因此，建立一定规模、可共享的生命科学专业学术英语语料库十分必要。

BioDEAP生命科学学术英语语料库是北京外国语大学中国外语与教育研究中心语料库共建项目DEAP学术英语语料库（Database of English for Academic Purposes）子库。BioDEAP语料库包含1,023个文本，共计5,131,276词次。

2. 建库目标

BioDEAP语料库的宗旨是服务我国语言教学及研究，创建一个覆盖我国生命科学领域12个二级学科领域权威英语期刊常用语类的全文数据库。设计总库容量为500万词。该语料库由研究论文与综述、讨论、书评、通讯四个学术语类子库构成，其中研究论文与综述子库下设12个分库（见表1）。

表1　生命学科学术英语语料库结构

1.研究论文与综述		2.讨论	3.书评	4.通讯
植物学	遗传学			
动物学	发育生物学			
生理学	细胞生物学			
水生生物学	生物化学与分子生物学			
微生物学	生物信息学与计算生物学			
神经生物学	生物物理学			

3. 语料收集方案

根据语料库建设分层取样原则，语料采集时我们遵循了以下4个原则：1）学科领域的代表性。按照国务院学位委员会、教育部2011年印发的学科专业目录生物学（学科代码0710）一级学科下设的植物学、动物学、生理学、水生生物学、微生物学、神经生物学、遗传学、发育生物学、细胞生物学、生物化学与分子生物学、生物信息学与计算生物学、生物物理学等12个二级学科分别选取期刊文章。2）期刊的权威性。以中科院2015年JCR（*Journal Citation Report*）收录的8,618种理工科学术期刊分区表为主要参考并咨询行业专家意见，每个二级学科选取5-6种期刊，最终共确定53本期刊（见表2）。3）语类的平衡性。按研究论文50%、综述文章35%、书评4%、讨论10%（包括评论、观点、论坛、访谈、科学与社会等栏目文章）、通讯1%（包括通讯、简讯、新闻报道）的比例抽选文章。文章类型的确定可以参照爱思唯尔（Elsevier）期刊库高级检索分类选项或者科学网文献类型选项。根据先导研究，每篇研究论文约6,500词，综述文章平均7,500词，书评约1,100词，讨论类文章约2,100词，通讯约1,500词，据此可以确定各子库的文本数量及抽取范围。4）年代的最新性。所有文本全文收录（包括主题词、摘要、致谢部分）发表时间尽量控制在2010—2017年期间。

表2　生命科学学术英语语料库来源期刊

序号	二级学科	刊物名称
1	Plant Science	*Plant, Cell and Environment*
2		*Plant Journal*

（待续）

（续表）

序号	二级学科	刊物名称
3	Plant Science	*Plant Cell*
4		*Trends in Plant Science*
5	Genetics	*Annual Review of Genetics*
6		*Nature Genetics*
7		*Nature Reviews Genetics*
8		*Trends in Genetics*
9	Cell biology	*Cell Stem Cell*
10		*Cell Research*
11		*Cancer Cell*
12		*Nature Cell Biology*
13		*Cell Metabolism*
14		*Cell Reports*
15	Bioinformatics	*Bioinformatics*
16		*Briefings in Bioinformatics*
17		*PNAS*
18		*PLoS Computational Biology*
19	Zoology	*Wildlife Monographs*
20		*Mammal Review*
21		*Animal behaviour*
22		*Zoologica Scripta*
23		*Behavioral Ecology and Sociobiology*
24		*Mammal Review*
25	Biophysics	*Annual Review of Biophysics*
26		*Nature Structural & Molecular Biology*
27		*Physics of Life Reviews*
28		*Physics of Life Reviews*
29	Deve biology	*Genes & Development*

（待续）

（续表）

序号	二级学科	刊物名称
30	Deve biology	*Developmental Cell*
31		*Annual Review of Cell and Developmental Biology*
32		*Seminars in Cell & Developmental Biology*
33	Physiology	*Annual Review of Physiology*
34		*Journal of General Physiology*
35		*Chronobiology International*
36	Neurobiology	*Nature Communications*
37		*Neuroscience and Biobehavioral Reviews*
38		*Experimental Neurology*
39		*International Journal of Neural Systems*
40		*Lancet*
41	Microbiology	*Cell Host & Microbe*
42		*Annual Review of Microbiology*
43		*Environmental Microbiology*
44		*Trends in Biotechnology*
45		*Nature Biotechnology*
46	Hydrabiology	*Marine Biology*
47		*Marine pollution bulletin*
48		*Harmful Algae*
49		*Freshwater biology*
50	Biochemistry	*Advanced Drug Delivery Reviews*
51		*Best Practice & Research Clinical Endocrinology & Metabolism*
52		*Soil Biology and Biochemistry*

4. 文本命名

为方便和DEAP语料库总库汇总，生命科学语料库文件命名采用“一级学科–二级学科–文献序号–文体”的顺序对所采集的文件进行分类命名。其中，一级

学科biology用首字母B表示；二级学科采用两个英文字母缩写（见表3）。文本语类有以下五种：1）研究论文research article（RA）；2）综述文章review article（RV）；3）书评book review（BR）；4）讨论类文章discussion（DC）；5）通讯correspondences（CP）。例如，文件名BBC082RA表明该文本是生命科学语料库中细胞生物学二级学科的第82篇文献，语类为研究论文。

表3　二级学科名称代码

二级学科	植物学	动物学	生理学	水生生物学	微生物学	神经生物学
代码	PS	ZL	PH	HB	MB	NB
二级学科	遗传学	发育生物学	细胞生物学	生物化学与分子生物学	生物信息学与计算生物学	生物物理学
代码	GE	DB	CB	BC	BI	BP

5. 文本转换与清理

首先，根据前期确定的文本采集框架、语料抽样原则分类下载文献，以PDF格式按照子学科体系编码分类存档，文献来源信息另外存档。之后用ABBYY FineReader 14将PDF文献批量转换为Word格式，再用DOC to TXT软件转为TXT文档保存。

第二步文本清理：人工逐一排查txt文档，对照PDF原文删除图表、公式、乱码、空格。清除页眉页脚信息（保留必要的元信息以便之后文本标注）。之后用Text Cleaning Library软件做噪音排除、断行修正及文本精确整理（见北外语料库语言学网站http://corpus.bfsu.edu.cn/channels/tools）。最后用AntConc软件wordlist - sort by - invert order功能检查是否有明显的拼写错误。如consistentlydresulted经查对原文应该是consistently-resulted。

6. 语料库标注

语料加工是语料库建设的关键环节，对于提升语料库的价值和语料库的后期应用至关重要。McEnery *et al.*（2006：22-45）将其分为标记（markup）和标注（annotation）。前者包括文本序号、文本分类、作者、来源、文本结构等；后者主要是指添加语言信息，如词性（POS）标注、语法标注、错误标注、语用赋码等（李文中 2012）。

6.1 表头信息

本研究采用标准XML格式，对表头信息及结构信息做了标记。内容包括文献序号、出版日期、学科领域、二级学科名称、语类、期刊名称、作者及所属机构、数字对象唯一标识符（DOI）、文章标题。书评类文本还对原书信息（书名、作者等）做了标记。结构信息包括摘要、引言、方法、结果、讨论、致谢等部分。以下是文档BHB697RA部分表头信息：

<Header>
<File_ID><BHB697RA></File_ID>
<Publication_Year>2018</Publication_Year>
<Domain>Science</Domain>
<Discipline>Biology</Discipline>
<Subdiscipline>Hydrobiology</Subdiscipline>
<Article_Type>Research Article</Article_Type>
……
</Header>

6.2 词性赋码

生命科学学术英语语料库利用自动词性标注软件TreeTagger（http://corpus.bfsu.edu.cn/tools）对语料做了词性标注。该软件由德国斯图加特大学Helmut Schmid开发，含有56个词性标记，正确率高达96%以上（Schmid 1994）。赋码语料库提供了更丰富的语言信息，为我们在词汇、句法、语法、语篇等多层次分析学术英语特点奠定了基础。

7. 生命科学语料库的应用展望

本语料库的创建有广阔的教学应用和语言研究前景。语料库分析软件的一个主要功能就是生成词表。Coxhead（2000）基于其自建的学术英语语料库推出了一份含570个词族的“学术词汇表”（Academic Word List）。Gardner & Davies（2014）采用不同的统计方法，发表了基于美国当代英语语料库学术子库的“新学术词汇表”（New Academic Vocabulary List），希望能够更广泛地涵盖学术语篇词汇。然而，许多学者（Hyland & Tse 2007；吴瑾、王同顺 2007；刘佳、韩丽娜 2014）对通用学术词表（例如AWL）的适用性提出质疑，这是因为基于语料库的研究表明，词汇的使用在范围、频率、搭配和意义等方面有很大差异。因此，Lei & Liu

（2016）认为有必要为各种特定学科制定学术词汇表。国内外学者用不同统计方法先后推出了石油、航海、护理、医学等专业词表（Yang 2015；Lei & Liu 2016；王京 2006；江淑娟 2010；赵志刚 2015），但是针对生命学科词表研究还很不充分。基于BioDEAP语料库创建生命专业学术英语词表，或创建细胞生物学、动物学、植物学等二级学科学术英语词表，无论从研究的角度或是教学的角度都非常必要。这些词表可用于词典编纂、专业英语教学大纲或教材编写，也可用于教学诊断等其他用途。学术语块的使用是衡量学习者语言能力的重要指标（Wray 2002；马蓉 2017），基于词表的常用学术英语词汇搭配表、词簇表有助于提高学生写作能力。

BioDEAP语料库涵盖生命学科多个领域、多种语类，专家学者可根据研究需要，利用Sub-corpus creator（许家金、梁茂成 2011，北京外国语大学语料库团队免费共享软件下载地址，http://corpus.bfsu.edu.cn/tools）自行创建子语料库做对比分析。研究对象既可以是不同母语背景的学者或学习者，也可以是跨语类、跨学科门类的学术语篇元话语。既可以在名词化、词的搭配、同义词辨析等词汇层面做深入探讨，也可以分析不同学科领域词汇、句法使用的差异。

8. 结语

桂诗春等（2010：420）指出“语料库语言学对资源有很大的依赖性，为了促进语料库语言学在我国的发展，应该提倡资源共享”。大型DEAP学术英语语料库的终极目标是超过一亿词次，作为其中的一个子库，BioDEAP生命科学学术英语语料库的创建正是秉承共享的理念。本项目在建设中严格按照DEAP语料库预先制定的统一规范，语料收集时以业内专家、教授意见为指导，精选学科权威期刊、进行文本清理、信息标注、语料入库等各环节，分工明确，双重检验，确保了语料的代表性、真实性以及标注的准确性。本项目成果将由北京外国语大学中国外语与教育研究中心语料库语言学团队统一发布在语料云网站或CQPweb网站，供语言研究者、语言教师、学生、科研人员免费使用。

参考文献

Coxhead, A. 2000. A new academic word list [J]. *TESOL Quarterly* 34(2): 213-238.

Gardner, D. & M. Davies. 2014. A new academic vocabulary list [J]. *Applied Linguistics* 35(3): 305-327.

Hyland, K. & P. Tse. 2007. Is there an academic vocabulary? [J]. *TESOL Quarterly* 41(2): 235-253.

Lee, D. Y. W. 2010. What corpora are available? [A]. In A. O’Keeffe & M. McCarthy (eds.). *The Routledge Handbook of Corpus Linguistics* [C]. London: Routledge. 107-121.

Lei, L. & D. Liu, 2016. A new medical academic word list: A corpus-based study with enhanced methodology [J]. *Journal of English for Academic Purposes* 22: 42-53.

McEnery, T., R. Xiao & Y. Tono. 2006. *Corpus-based Language Studies: An Advanced Resource*

Book [M]. London: Routledge.

Schmid, H. 1994. Probabilistic part-of-speech tagging using decision trees [A]. In *Proceedings of International Conference on New Methods in Language Processing,* Manchester, UK [C]. Retrived from: http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/

Wray, A. 2002. *Formulaic Language and the Lexicon* [M]. Cambridge: CUP.

Yang, M. 2015. A nursing academic word list [J]. *English for Specific Purposes* 37: 27-38.

范晶晶、李丽霞，2014，农业学术英语语料库建设构想[J],《安徽农业科学》(7)：2169-2170。

桂诗春、冯志伟、杨惠中、何安平、卫乃兴、李文中、梁茂成，2010，语料库语言学与中国外语教学[J],《现代外语》(4)：419-426。

何中清、彭宣维，2011，英语语料库研究综述：回顾、现状与展望[J],《外语教学》(1)：6-10，15。

江淑娟，2010，石油英语学术词汇表创建研究[J],《西南石油大学学报（社会科学版）》(6)：27-30。

李丽霞、白　璐，2016，基于VOA农业新闻英语语料库的主题词分析[J],《安徽农业科学》(6)：330-332。

李文中，2012，语料库标记与标注：以中国英语语料库为例[J],《外语教学与研究》(3)：336-345。

梁茂成、李文中、许家金，2010,《语料库应用教程》[M]。北京：外语教学与研究出版社。

刘　佳、韩丽娜，2014，基于语料库的通用学术词表在专业英语学习中的适用性研究——以“学术词表”和环境科学专业为例[J],《北京化工大学学报（社会科学版）》(1)：63，64-68。

刘　萍、黄小倩、刘　珊，2015，农科学术英语论文语料库的创建[J],《语料库语言学》(2)：97-106。

马　蓉，2017，学术词汇研究四十五年[J],《现代外语》(3)：420-428。

陶　玲，2017，建立小型专业化语料库 完善农业科技期刊英文摘要编写[J],《吉林农业》(4)：107。

王　京，2006，基于医学研究论文语料库的医学学术词表的构建[D]。第四军医大学硕士学位论文。

王　敏、李丽霞，2014，动物科学国际期刊论文语料库的创建与应用[J],《安徽农业科学》(20)：6854-6856。

吴　瑾、王同顺，2007，Coxhead“学术词汇表”的适用性研究[J],《国外外语教学》(2)：28-33。

吴　蕾、赵晓临、张继东，2014，专业科技英语语料库的建设与应用[J],《东华大学学报（社会科学版）》(2)：81-85。

许家金、梁茂成，2011，创建子语料库，促成对比研究[J],《当代外语研究》(10)：6-9。

张新杰，2017，国内语料库语言学研究：回顾与展望——基于核心期刊24年文献的统计分析[J],《西安外国语大学学报》(2)：36-41。

赵志刚，2015，专门用途英语学术词表创建研究——以航海英语为例[J],《重庆交通大学学报（社会科学版）》(6)：140-144。

通讯地址：100049 北京市中国科学院大学

LinDEAP语言学学术英语语料库的创建*

青岛大学 布占廷 王 昕 王 乐

提要：语言学学术英语语料库是DEAP学术英语语料库的一个重要组成部分。本文介绍LinDEAP语言学学术英语语料库的建库目标、取样考量、实施方案和应用前景等内容。首先，本文论述语料取样的代表性，主要涉及学科覆盖面、期刊来源、文本年份、文本语类、作者母语等方面的考量。其次，详细说明LinDEAP语料库的文本组织与命名、文本清理与标注等。最后，探讨了LinDEAP语料库的发布及其在语言学学术英语教学与科研领域的应用前景。

关键词：LinDEAP、语言学学术英语、语料库

1. 引言

语料库是“自然发生的语言数据的集合，可作为语言研究的基础”（Sinclair 1991：171）。通常是在一定的采样原则下收集一定规模的真实语料，以电子文本形式存储于计算机之中，用于语言的量化研究和质性分析，这些语料可以代表某种语言、某种语言变体或语类。因此，语料库有着不同的类型，从应用层面可以分为通用语料库和专用语料库。专用语料库又称为专题语料库，往往只收集某个特定领域或主题的语料样本，能够满足特定的研究目的，也能为编制专门领域的参考资料提供理想材料。因此，专用语料库又可以根据不同使用目的进行细分。

1982年，上海交通大学科技英语语料库（JDEST）启动建设。迄今为止，国际上已经开发了很多ESP专用语料库。比较著名的有Ken Hyland建设的多学科学术期刊论文数据库（8个学科领域，240篇期刊论文，130多万词次），John Swales开发的学术口语语料库（Michigan Corpus of Academic Spoken English，录音约200小时，170万词次），英国考文垂、雷丁等大学联合建设的英国学术英语写作语料库BAWE（British Academic Written English）等。20世纪90年代以来，专门用途语料库在我国发展迅猛，很多学科都相继建立了语料库（徐秀玲、许家金 2017）。

具体到语言学学科领域，桂诗春（2009）建立了一个100万词的英语语言学语料库ECOL（English Corpus of Linguistics），其中语料覆盖十大分支学科。在此

* 本文系第八批中国外语教育基金项目“学术英语语料库子库（ZGWYJYJJ2016A02）”的阶段性成果。

基础上，他用语料库的方法归纳出语言学语体的一些特征。桂诗春（2014）提到他还建设过一个我国硕士和博士研究生语言学论文数据库，库容约为50万词次，其目的是以ECOL为参照，研究硕、博士论文写作的特点和问题。但因语料的代表性和论文写作规范等方面存在问题，该语料库并未公开。另外，其他学者也建立了一些小型的语言学学术英语语料库，一般涵盖语言学中一个或者少数几个分支学科，用于特定的研究目的。例如，郑红红（2014）以《中国应用语言学》期刊2005年1月至2010年6月间发表的中国学者撰写的英文论文为语料，建设了中国学者学术英语语料库CAWEC（Chinese Academic Written English Corpus），用以研究中国学者应用语言学英语论文中的词块。在此项研究中，她以马晓雷建设的国外学者学术英语语料库FAWEC（Foreign Academic Written English Corpus）作为参照进行对比分析，语料源于*TESOL Quarterly*、*Language Learning*、*Applied Linguistics*和*Studies in Second Language Acquisition*。梁茂成、刘霞（2014）从国际期刊*Applied Linguistics*中抽取170篇学术论文建设专用语料库，并以此为例，借助TextSmith Tools工具分析了学术论文各部分的短语学特征。杨林秀（2015）建设了一个语言学学术英语语料库，包含100篇学术论文，主要来自*Journal of English for Academic Purposes*（2004-2008），*English for Specific Purposes*（2004-2008）和*Journal of Pragmatics*（2004-2008）等期刊，她从言据性视角考察了英语学术论文中作者身份的建构问题。王芙蓉、王宏俐（2015）建设了语言学和工科学术期刊论文语料库，并对两个库中的四词词块的结构固定性、结构形式特点、语篇功能等进行了对比分析。这些语料库的建成和使用，有力促进了语言学学术英语的研究与教学应用。但是，目前仍然存在着资源难以共享、语料复用性不强、发展不够平衡、建库规范和语料选择标准不一致等问题。另外，部分语料库由于未能充分考虑到建库原则以及标注方法等，建成的语料库存在一些缺陷。

以下将介绍LinDEAP语言学学术英语语料库的建库流程。LinDEAP是DEAP学术英语语料库（Database of English for Academic Purposes）的语言学子库，建成后包含626个文本，共计5,291,806个形符。LinDEAP在“中国外语教育基金专用英语语料库建设项目”的框架下设计和建设。因此，建库标准和建库流程与其他领域，如化学、法学、哲学等子库基本一致，这为不同领域语料库间的对比研究提供了基础和平台。本研究主要介绍语言学学术英语语料库的建库目标、建库过程和应用前景。

2. 建库目标

LinDEAP的建设目标包括两方面。其一，结合语言学的学科特点，建成一个涵盖面广、规模较大、时效性强、能体现语言学学术英语语言特征的专用语料库，探索LinDEAP规范的建库原则和标注方法。LinDEAP涵盖语言学主要分支学科，

设计库容为500万词次，语料来源为语言学国际SSCI检索高影响因子期刊在2014-2016年间发表的论文。语料库经清理后，采用标准XML编码格式进行标注。其二，服务于语言学学术英语的教学和科研，为术语提取、教材编写、语言学学术英语研究等提供真实的语言数据支持。

3. 文本收集方案

3.1 学科

建设英语语言学语料库，首先要考虑语料的覆盖面，从而“达到以一定大小的语言样本代表某一研究所确定的语言运用总体”（杨惠中 2002：9）。我们基于语言学一级学科，参照《中华人民共和国学科分类与代码简表（国家标准GBT 13745-2009）》（以下简称为《国标》）和桂诗春（2009）对语言学分支学科的分类，设立十个分支学科子库。

依据《国标》（见表1），语言学分为十个分支学科。但这种分类很难直接作为我们语言学语料库学科分类的依据。其中一个主要原因是，“汉语研究”、“中国少数民族语言文字”和“外国语言”是依据语种划分的，与其他七个类别划分标准明显不同，且很大程度上相互交叉。另一方面，国际语言学期刊论文研究内容所涉语种以英语为主，针对汉语的研究较少，而针对中国少数民族语言文字的研究更少。因此，以此作为学科分类依据，不同子库存在交叉较大，子库与期刊较难匹配，子库语料的平衡也难有保障。

表1 《国标》中语言学的分支学科

740	语言学	说明
74010	普通语言学	语音学；语法学；语义学；词汇学；语用学；方言学；修辞学；文字学；语源学；普通语言学其他学科
74015	比较语言学	历史比较语言学；类型比较语言学；双语对比语言学；比较语言学其他学科
74020	语言地理学	
74025	社会语言学	
74030	心理语言学	
74035	应用语言学	语言教学；话语语言学；实验语音学；数理语言学；计算语言学；翻译学；术语学；应用语言学其他学科

（待续）

（续表）

740	语言学	说明
74040	汉语研究	普通话；汉语方言；汉语语音；汉语音韵；汉语语法；汉语词汇；汉语训诂；汉语修辞；汉字规范；汉语史；汉语研究其他学科
74045	中国少数民族语言文字	蒙古语文；藏语文；维吾尔语文；哈萨克语文；满语文；朝鲜语文；傣族语文；彝族语文；壮语文；苗语文；瑶语文；柯尔克孜语文；锡伯语文；中国少数民族语言文字其他学科
74050	外国语言	英语；德语；瑞典语；丹麦语、挪威语、冰岛语；拉丁语；意大利语；法语；外国语言其他学科
74099	语言学其他学科	

桂诗春（2009）建设的英语语言学语料库ECOL覆盖了语言学的各个学科，分为十个子库（见表 2）。每个子库涵盖面不同，样本数也不同，共有500篇，每篇约2000词，语料选自语言学专著、教科书和学术期刊。

表2　ECOL的样本分布（桂诗春 2009: 20）

学科分类	篇数
1. Al（应用语言学、二语习得、语言测试等）	100
2. Cg（认知语言学、认知科学等）	38
3. Co（语料库语言学）	36
4. L（自然语言处理、神经语言学、统计语言学、生物语言学等）	70
5. P（心理语言学）	38
6. Pr（语用学）	38
7. Se（语义学、词汇学等）	36
8. So（社会语言学、文化与语言等）	36
9. St（文体学、语篇分析、翻译学等）	38
10. T（理论语言学、普通语言学、语法学、历史语言学、比较语言学等）	70
总计	500

尽管桂诗春（2009：20）坦言“这不是科学的分类”，但这一分类为语言学语料库建设中的分支学科分类问题做出了有益的探索。LinDEAP以桂诗春（2009）对语言学分支学科的划分为主要参照，同样设定十个子库，并赋以名称、编号与英语缩写（见表3），以便于标注与指称。

表3　分支学科名称表

分支学科		期刊		
名称（编号+缩写）	样本数量	名称	缩写	样本数量
应用语言学（01AL）	56	*Applied Linguistics*	AL	16
		Computer Assisted Language Learning	CALL	15
		Language Learning	LL	8
		Language Teaching	LTea	8
		Studies in Second Language Acquisition	SSLA	9
认知语言学（02CG）	59	*Cognitive Linguistics*	CL	21
		Language Cognition and Neuroscience	LCN	20
		Metaphor and Symbol	MS	18
语料库语言学（03CO）	55	*Corpus Linguistics and Linguistic Theory*	CLLT	20
		International Journal of Corpus Linguistics	IJCL	35
自然语言处理（04NL）	52	*Computational Linguistics*	CmL	20
		Journal of Quantitative Linguistics	JQL	12
		Natural Language Engineering	NLE	20
心理语言学（05PL）	53	*Applied Psycholinguistics*	AP	12
		Brain and Language	BL	20
		Journal of Memory and Language	JML	13
		Journal of Neurolinguistics	JN	8

（待续）

（续表）

分支学科		期刊		
名称（编号+缩写）	样本数量	名称	缩写	样本数量
语用学（06PR）	81	*Journal of Pragmatics*	JoP	44
		Linguistics and Philosophy	LP	21
		Pragmatics	PRAG	16
语义学与词汇学（07SL）	55	*International Journal of Lexicography*	IJL	15
		Journal of Semantics	JSe	25
		Natural Language Semantics	NLS	15
社会语言学（08SO）	91	*Journal of Sociolinguistics*	JoS	30
		Language in Society	LiS	36
		Research on Language and Social Interaction	RLSI	25
语篇分析和文体学（09DS）	55	*Journal of Fluency Disorders*	JFD	16
		Language & Communication	LC	26
		Text & Talk	TT	13
理论语言学（10TL）	69	*Journal of Linguistics*	JL	9
		Language	LANG	24
		Language and Speech	LnS	16
		Language Sciences	LS	12
		Linguistic Inquiry	LI	8
合计	626			626

需要注意的是，我们在桂诗春（2009）的基础上对每个分支学科的覆盖范围做了一些微调。具体而言，01AL主要包括应用语言学、二语习得、语言测试、语言教学和术语学等。02CG主要包括认知语言学和认知科学等。03CO主要包括语料库语言学。04NL主要包括自然语言处理、统计语言学、生物语言学、计算语言学、数学语言学[1]、量化语言学和生态语言学等。05PL主要包括心理语言学、实验语言学[2]和神经语言学等。06PR主要包括语用学[3]和语言哲学等。07SL主要包括语义学[4]、词汇学和词源学等。08SO主要包括社会语言学、文化与语言、法律语言学和社会符号学等。09DS主要包括语篇分析、文体学、翻译学、手语、语言

人类学和言语民族志等。10TL主要包括理论语言学、普通语言学、语法学、句法学、语音学、音系学和语文学、比较语言学、对比语言学等。

3.2 期刊

以学科分类为参照，我们对*Journal Citation Reports*（2016）中的语言学期刊目录进行了分析。该报告按照前一年（即2015）期刊的影响因子，罗列了177种语言学期刊。这些期刊可以归入表3中的十个子库。我们以此为依据在每个子库选择2-5个高影响因子期刊作为语料来源[5]，共选择期刊34个，并设定了期刊名称缩写。缩写一般是采用期刊名称中实义单词首字母组合而成。但有几种例外情况。首先，期刊名称只有一个单词，如*Language*和*Pragmatics*，其缩写设定为单词的前四个字母。其次，为了避免重复，在缩写中加入了小写字母，如期刊*Computational Linguistics*的缩写设定为CmL，以便与期刊*Cognitive Linguistics*（CL）区分开来。这种区分也能为后续语料库扩容做好铺垫。再次，某些刊物本身固有的缩写中含有小写字母等区别性元素，如*Journal of Pragmatics*（JoP）。

3.3 时间

我们将语料收集的时间限定在2014年到2016年，时间跨度较小，能够更好地反映语料的时代性和当下语言学学术英语的语言特征。

3.4 语类

我们共收集了三种语类的论文，分别是：1）研究性论文（research articles），缩写为A，是语言学学术研究中最具代表性的论文类型。几乎所有的语言学期刊均刊登此类论文，报道本领域最新研究成果。2）综述论文（review articles），缩写为RA，是指某一特定领域研究历史、进展与动态的综述性研究，此类文章只出现在少量的语言学期刊上。3）书评（book review），缩写为BR，是指对最近出版的一本或一本以上的学术著作的介绍和评论性文章，此类文章只出现在部分语言学期刊上。每篇文章都是全文收录，主要包括标题、摘要、关键词、正文和致谢等。

3.5 作者母语

原则上要有一位作者是英语母语者（依据姓名、国籍、工作单位、作者介绍等信息来判断）。大部分文本的作者来自美国和英国，少部分来自加拿大、澳大利亚和新西兰等英语国家。

基于以上考量，我们进行语料收集和下载工作。这一工作集中在2017年8月-11月完成，下载渠道是学校购买的数据库资源。语料正文下载分为两种方式：

1）如选中的文本全文有HTML格式，则直接拷贝后粘贴到设定好的空白Microsoft Word文档中，同时也下载相应的PDF版，以备后用。这种方法的优点有：断行和乱码现象很少；操作简单易行。2）如选中的文本全文没有HTML版，则直接下载PDF版。通过第一种方式下载的部分语料中会有view image等指示语，需要清理。通过第二种方式下载的语料需要转换格式，并做文本清理。

在下载语料的同时，我们也下载了语料的引用信息（citation），直接导入到文献管理软件EndNote中，用于元信息整理。这些信息最后导出为Excel文件，其中主要包括作者、发表年代、论文题目、期刊名称、卷号、期号、起止页码、ISSN编号、DOI号和URL。在此基础上进一步整理，添加文本编号和子库名称，即可形成最终的Excel版元信息文件。语料数量见表3。共有626个完整文本，其中研究性论文522篇，综述论文22篇，书评82篇，总库容为5,291,806词次。

4. 文本的组织与命名

文本存放下载目录直接按照期刊名设置文件夹。文件夹名称命名规则为“子库编号-子库缩写-期刊名缩写-期刊名全称”，如文件夹01 AL CALL Computer Assisted Language Learning中存放源自期刊*Computer Assisted Language Learning*的语料。这种扁平化的组织形式易于管理，便于排序查看。

文件名称命名规则为“子库编号–缩写–子库内编号–期刊缩写–语类”，此类命名具有较高的自足性。例如，在01AL01-AL-A这个名称中，第一个01表示子库编号，AL表示子库名称缩写，第二个01表示该子库的第一篇语料，AL表示期刊名称缩写，A表示语料类型是研究论文。

5. 文本的清理与标注

文本的清理与标注工作是语料库建设工作的关键环节，全部由人工完成，耗时费力，包含研究者在内一共有七名工作人员。在工作开始前，研究者对工作人员进行了培训，确保所有人熟悉文本清理与标注的内容与标准。同时，研究者分发详细流程，便于自我对照检查。另外，清理与标注完成后，研究者对所有文本进行了核查，在整个过程中，要及时做好文本备份工作。研究者与标注者建立协同机制，始终保持联系，对存在的疑惑和歧义进行讨论，提高结果的准确性和一致性，保证后续研究的效度与信度。

5.1 文本清理

文本整理主要涉及文本提取、降噪等。首先是文本提取。通过第二种方式下载的语料为PDF格式文件，需要首先转换为Word文件。转换前，使用软件Adobe

Acrobat Pro对论文页面进行裁剪，去除页眉和页脚。然后通过ABBYY FineReader将PDF文件转换为Microsoft Word文件。

其次是文本降噪。通过网络下载和识别转换等渠道获得的语料存在各种不合规范的符号、格式、乱码等等，会导致后续的检索与标注出错，因此需要进行降噪处理。这一过程主要是在Word环境下进行的。首先，打开Word文件，比照PDF原文，逐一清理转换过程中出现的标点错误、乱码、板块结构顺序错乱、段落切断、拼写错误、项目编号和例子编号错乱等问题，并对语料进行校对。在此过程中仔细观察语料，发现规律，使用Word的宏功能进行批量处理，提高处理效率。其次，删除文本中无法体现语言特征的数据，如论文封面、图表（包括相应标题与注释）、公式，以及脚注、尾注、附录、项目资助信息、参考文献、论文声明（discoursal statement）、附加数据（supplemental data）等语言学信息不强的部分。需要说明的是，部分公式出现在行文中，公式的删减对行文的完整性造成了一些影响。

5.2 文本标注

标注能够为文本增加额外信息，有助于开展后续研究，从而为语料增值（Leech 1991），这也已经成为大型语料库最重要的基本规范之一。本研究采用标准的XML编码格式对语料库中的元信息和文本结构信息进行标注，从而为语料库检索和分析提供条件和依据。XML标注成对出现，如<Discipline></Discipline>，其中前者是开始标记，后者是关闭标记。标注开始前，需要组织专门的培训，详细讲解，确保标注的准确性和一致性。XML标注结果有很好的兼容性，能被不同应用程序解析和使用（邢富坤 2015）。

元信息是指为文本提供的更多额外信息，对语料库研究具有非常重要的意义，可以提供子学科分类、发表期刊、作者等方面的信息，据此可以生成不同的子库。元信息标注在每篇文本的开始位置。标注信息与释义参见表4。对于个别缺失而且无法找到的信息，标注为unknown。

表4 元信息标注

元信息标注符	释义
<Header>	头部信息开始
<Discipline></Discipline>	学科
<Domain></Domain>	分支学科子库
<References>	文献信息开始

（待续）

（续表）

元信息标注符	释义
<Article_Title></Article_Title>	文章题目
<Journal_Title></Journal_Title>	期刊名称
<Article_Type></Article_Type>	文章类型
<Volume></Volume>	卷号
<Issue></Issue>	期号
<Pages></Pages>	页码
<Publisher></Publisher>	出版社
<DOI></DOI>	DOI码
<Publication_Year></Publication_Year>	发表年代
<Author></Author>	作者
<Affiliation></Affiliation>	作者单位
</References>	文献信息关闭
</Header>	头部信息关闭

元信息中作者（Author）与作者单位（Affiliation）可能会涉及两个及两个以上的人和机构，原文一般采用阿拉伯数字、英文小写字母、缺省等编码方式，本研究中将其统一为阿拉伯数字编码方式。如文本01AL01-AL-A的作者和作者单位信息标注如下：

<Author>1, Tess Fitzpatrick; 2, David Playfoot; 1, Alison Wray; 3, Margaret J. Wright</Author>

<Affiliation>1, Centre for Language and Communication Research, Cardiff University, UK; 2, Department of Psychology, Sociology and Politics, Sheffield Hallam University, UK; 3, Genetics and Computational Biology, Queensland Institute of Medical Research, Brisbane, Australia</Affiliation>

文本内容标注是指对文本的结构进行XML标注。这一标注始于论文摘要，止于文章结尾。这一标注为我们进行语类分析、子库生成等提供了基础条件。我们对三种语类做了区分对待。首先，研究论文的标注信息参见表5。标注的层级局限于一级标题，没有深入到二级、三级标题，否则标注信息会出现较多的重叠与嵌套现象，造成混乱，也不利于后续研究的开展。在标注过程中，我们基于文本结构的实际情况，从表 5中选择合适的标注符号进行标注。其次，综述论文正文的

主体部分主要是文献综述（Literature Review），其前一般会有引言（Introduction），其后往往会有结论（Conclusion）。再次，没有对书评正文的内部结构做进一步的标注。

表5　文本结构标注信息

文本结构标注信息	意义
<Body>	正文开始标记
<Abstract></Abstract>	摘要
<Keywords></Keywords>	关键词
<Introduction></Introduction>	引言
<Literature_Review></Literature_Review>	文献综述
<Research_Design></Research_Design>	研究设计
<Methods></Methods>	方法
<Results></Results>	结果
<Discussion></Discussion>	讨论
<Results_Discussion></Results_Discussion>	结果与讨论
<Experiment></Experiment>	实验
<Experimental_Procedures ></Experimental_Procedures >	实验程序
<Discussion_Conclusion></Discussion_Conclusion>	讨论与结论
<Conclusion></Conclusion>	结论
<Limitations></Limitations>	局限性
<Limitations_Future Research></Limitations_Future Research>	局限与将来研究
<Acknowledgments></Acknowledgments>	致谢
</Body>	正文关闭标记

在所有清理和标注工作完成之后，研究者利用“DOC to TXT”软件将Word文本批量转换为TXT文本，并进行检查，然后依据文本组织进行存放，形成语料库。

6. 基于LinDEAP语料库的教学应用和语言研究

建成后的LinDEAP语言学学术英语语料库将发布在语料云网站或CQPweb网站，供英语学习者、外语教师以及语言学教学与科研人员通过浏览器进行在线检索，减轻计算机终端存储负担，方便在网络环境下随时调用，同时也能避免版权等问题的困扰，实现资源共享，有效服务于语言学学术英语教学与研究（参见许家金、吴良平 2014）。

LinDEAP可应用于教学，为教学提供更有针对性的语言素材。语料全部来自于国际语言学领域的高影响因子期刊，语料真实，具有示范性，应用于教学实践，能够提升教学效果。首先，LinDEAP可以直接作为教学资源应用于教学，如语言学类课程，尤其是语料库语言学课程和学术论文写作课程，如“语料库语言学”、“如何利用语料库”、“用语料库开展教学”。其次，LinDEAP也可以间接地应用在教学中，如用于编写教材、大纲词表、测试材料等。

在基于语料库的语言学语体研究方面，首先，LinDEAP可以用于基本统计分析、语法特征分析和词汇特征分析，然后可以将结果与桂诗春（2009）进行对比。需要注意的是，桂诗春（2009）建设的英语语言学语料库包括500篇语料，每篇2, 000词，均不是完整语篇，语料来自语言学专著、教科书和学术期刊，语类多样且差异大。其次，可以系统考察语言学语篇的语类结构。我们在文本结构标注中发现，语步的数量和种类与语类、分支学科等变量之间存在着一定的关系，值得进一步深入研究。再次，LinDEAP可以用来生成若干个子库，如不同学科子库、不同语类子库，对多种语言特征进行对比分析。在此类研究中，LinDEAP也可以作为参照语料库，对某个语类子库中的语言特征进行主题词分析，发现过多或过少使用的语言项目。

LinDEAP还可以广泛地应用于语言学学科专用词表和词块表构建、词典编纂、机器辅助翻译、自然语言处理、术语学等方面的研究，也可以作为参照语料库应用于对其他英语语言学语料库的研究中。

7. 结语

语料库与学术英语研究之间存在着紧密的联系。LinDEAP团队按照既定建设目标，从学科覆盖面、期刊范围、文本年份、语类、作者母语等多方面对建库设计进行了认真规划，收集下载了2014-2016年间发表于国际34个语言学高影响因子刊物上的文章626篇，遵循一定的规则对文本进行有效组织与命名，严格按照规范对文本进行整理，多位标注者同时标注以提高标注的准确度。建成的LinDEAP语言学学术英语语料库涵盖面广，规模较大，具有很强的时效性，能够体现语言学学术英语的特征，为相关领域的教学科研提供有代表性的真实语言材

料。LinDEAP也具有开放性，可以根据研究需要进一步扩充与更新。

注　释

1. 在《国标》中，数学语言学被归入应用语言学。
2. 在《国标》中，实验语言学被归入应用语言学。
3. 在《国标》中，语用学被归入普通语言学。
4. 在《国标》中，词汇学、语义学和词源学等被归入普通语言学。
5. 需要注意的是，*Journal Citation Reports*（2016）列表中高影响因子刊物的学科分布是很不均衡的。例如，在前20名刊物中，应用语言学有10个，心理语言学有4个，二者之和接近总数的四分之三。

参考文献

Leech, G. 1991. The state of the art in corpus linguistics [A]. In K. Aijmer & B. Altenberg (eds.). *English Corpus Linguistics* [C]. London: Longman. 8-29.

Sinclair, J. 1991. *Corpus, Concordance, Collocation* [M]. Oxford: OUP.

桂诗春，2009，《基于语料库的英语语言学语体分析》[M]。北京：外语教学与研究出版社。

桂诗春，2014，语料库语言学答客问[J]，《语料库语言学》(1)：1-15。

梁茂成、刘　霞，2014，语篇内部的短语学特征分布模式探索——以学术论文为例[J]，《解放军外国语学院学报》(4)：1-11，22。

王芙蓉、王宏俐，2015，基于语料库的语言学和工科学术英语词块比较研究[J]，《外语界》(2)：16-24。

邢富坤，2015，面向语言处理的语料库标准：回顾与反思[J]，《解放军外国语学院学报》(3)：8-13。

许家金、吴良平，2014，基于网络的第四代语料库分析工具CQPweb及应用实例[J]，《外语电化教学》(5)：10-15，56。

徐秀玲、许家金，2017，我国外语教学中的语料库应用40年[J]，《中国外语教育》(4)：62-68。

杨惠中，2002，《语料库语言学导论》[M]。上海：上海外语教育出版社。

杨林秀，2015，英文学术论文中的作者身份构建：言据性视角[J]，《外语教学》(2)：21-25。

郑红红，2014，中国学者应用语言学英语论文中的词块研究[J]，《语料库语言学》(1)：47-57。

通讯地址：266071 山东省青岛市青岛大学外语学院

MilDEAP军事学术英语语料库的创建

国防科技大学 马晓雷 陈钻钻 张皓盟

提要：军事学术英语语料库是“DEAP学术英语语料库”的子库。本文从语料来源、语料收集、语料标注、语料入库等方面介绍该语料库的建设过程，并对其后续建设方案和应用前景进行探讨。最后，结合军队信息化建设，对军事语料库的发展前景提出建议。

关键词：军事学术英语、语料库、信息化建设

1. 引言

作为大规模机器可读的真实语言集合（常宝宝、俞士汶 2009），语料库不仅是文本库，同时也是知识库和规则库，更是各种语言信息处理研究必不可少的训练库和测试库。正因如此，语料库在语言研究、语言教学、自然语言处理等领域发挥着越来越重要的作用。然而近年来，基于语料库的应用型研究投入相对较多，也取得了一定的成果，但是在语料库这一基础资源的建设方面却相对进展缓慢。现有的语料库虽然规模逐渐扩大，但文本采样的代表性与平衡性问题大都没有得到很好的解决。此外，语料库的建库流程和标准不统一，导致相关主题的语料库之间难以兼容整合，造成资源的浪费。再者，语料库的加工深度仍然不足，尤其缺乏基于语言学理论的文本标注。

“中国外语教育基金专用英语语料库建设项目”旨在建设包含人文社会科学、自然科学各主要学科领域的总容量不少于一亿词的学术英语平衡语料库。相对于大规模通用语料库，面向特殊用途领域的专门语料库具有较强的针对性和实用性。由于范围相对明确，专门语料库建设在文本采样、流程标准、语料标注等方面更易于操作。作为子项目之一，“MilDEAP军事学术英语语料库”以总项目规定的建库原则为指导，采用标准的文本采样和语料整理流程，最终建成具有一定代表性和平衡性，且与医学、法律等子库相对应的500万词级的专用语料库。以往国内军事领域的语料库建设，主要聚焦军事翻译（王岚、严灿勋 2015）、军事后勤（李永芹等 2014）、军事医学（肖健等 2017）、军事装备（杨阳 2011）等领域。本课题以军事学术英语为文本采样对象，能够在一定程度上填补目前国内该领域的空白。同时，作为“中国外语教育基金专用英语语料库建设项目”的子课题，该项目在采样、清理、标注等方面采用统一标准，易于跨语料库开展对比分析，也有利于

军事类语料库的推广应用。

2. 语料库概况

MilDEAP军事学术英语语料库共收录24种国际军事学期刊发表的英语学术论文702篇，总词次为5,053,446，单篇论文长度约7,199词，标准形符类符比为40.88。语料库覆盖了军事医学、军事历史、军事法律、国际维和、反恐、航空工程、海洋工程、冲突管理和战略研究等军事领域，可以较好地代表军事学术英语的总体情况。语料库采用标准XML格式标注元信息，能够较好地与总项目中的其他子库相兼容。

3. 语料来源

不同于医学、法律、国际关系等学科领域，国际知名期刊数据库大都没有对军事学期刊进行单独归类。这使得课题组无法直接从Web of Science、EBSCO、SAGE等主流数据库中检索并下载军事学期刊论文。为解决这一问题，我们借鉴王涛等（2015）确定的军事学期刊列表，并结合Web of Science数据库的收录时间和论文总量，最终选定24种国际军事学期刊作为语料搜集来源（见表1）。

表1　国际军事学期刊列表

序号	期刊名称	Web of Science学科分类
1	*Journal of the Royal Army Medical Corps*	Medicine General Internal
2	*Aircraft Engineering and Aerospace Technology*	Engineering Aerospace
3	*Defence Science Journal*	Multidisciplinary Sciences
4	*Naval Engineers Journal*	Engineering Marine
5	*International Journal of Naval Architecture and Ocean Engineering*	Engineering Marine
6	*International Security*	International Relations
7	*Journal of Peace Research*	International Relations
8	*Conflict Management and Peace Science*	International Relations
9	*Survival*	International Relations
10	*Washington Quarterly*	International Relations

（待续）

（续表）

序号	期刊名称	Web of Science学科分类
11	*Terrorism and Political Violence*	International Relations
12	*Journal of Strategic Studies*	International Relations
13	*International Peacekeeping*	International Relations
14	*Intelligence*	Psychology Multidisciplinary
15	*Military Psychology*	Psychology Multidisciplinary
16	*Journal of Safety Research*	Ergonomics
17	*Studies in Conflict and Terrorism*	Social Sciences
18	*Armed Forces and Society*	Social Sciences
19	*Korean Journal of Defense Analysis*	Social Sciences
20	*Military Law Review*	Law
21	*Defence and Peace Economics*	Economics
22	*War in History*	History
23	*War & Society*	History
24	*Journal of Cold War Studies*	History

仅从名称上看，表1中的绝大部分期刊都明显与军事主题相关，如*Defence Science Journal*、*Military Psychology*、*Armed Forces and Society*、*Military Law Review*、*War in History*等。但也有少部分期刊的名称不包含与军事主题相关的关键词，如*Washington Quarterly*、*Aircraft Engineering and Aerospace Technology*，等等。从Web of Science提供的主题标签来看，以上24种期刊分属医学、工程、国际关系、心理、社会、法律、经济、人体工学、历史等领域，这某种程度上说明军事学研究具有一定的跨学科属性。总的来说，以这24种期刊为数据采集来源，具有一定的代表性。

4. 语料收集与清理

在确定目标期刊列表后，课题组通过Web of Science数据库检索出各个期刊被引用频次最高的50篇文献作为备选对象，并对其中的30篇进行下载。选取50篇备选文献的目的在于确保每种期刊均能下载30篇文献，避免出现部分文献下载链接失效或下载困难导致文本采样总量不足的问题。

下载的初始语料均为PDF格式，研究者使用ABBYY FineReader软件将PDF

文件批量转换为Word格式。该软件的优点在于转换后的Word文档在版面布局上与原PDF文件一致，便于后期对照检查。接着，课题组对文本进行清理，主要工作包括：删除论文封面、页眉页脚、图表公式、脚注尾注、分节符，纠正拼写错误与乱码、合并跨页段落等。图表删除后统一插入符号<T>，以保留原始文档中的图表位置信息。清理工作的主要目的是使文本成为干净的纯文本，便于后期进行文本标注。

5. 文本标注

文本标注的内容主要有元信息和文本结构信息（见表2和表3）。元信息出现在每一篇文本的开头，包括语料库名称、版本、建库单位、创建者姓名、文章标题、期刊名、发表时间、作者等。按项目总要求，全部标注均采用XML格式。

表2　元信息对照表

<CORPUS_NAME> </CORPUS_NAME>	语料库名称
<YLKBB> </YLKBB>	语料库版本
<JKDW> </JKDW>	语料库建库单位
<JKZZ> </JKZZ>	语料库作者
<JKSJ> </JKSJ>	语料库建库时间
<YLQLR> </YLQLR>	语料清理人
<ARTICLE_TITLE> </ARTICLE_TITLE>	文章标题
<JOURNAL_TITLE> </JOURNAL_TITLE>	期刊名称
<VOLUME> </VOLUME>	卷号
<ISSUE> </ISSUE>	期号
<PAGES> </PAGES>	页码
<DOI> </DOI>	DOI号
<PUBLICATION_YEAR> </PUBLICATION_YEAR>	出版年份
<AUTHOR> </AUTHOR>	作者

表3　文本结构信息对照表

<A> </A>	摘要
<H> </H>	标题（包括introduction、method、result、discussion等）
<R> </R>	参考文献（后期已删除）
<F> </F>	脚注
<Q> </Q>	直接引用
<L> </L>	列表信息
<E> </E>	例子
<N> </N>	尾注

文本标注工作由三位项目成员合作完成。在正式标注前，项目负责人结合预标注结果制定了详细的标注规范，并对三位标注者进行培训，确保他们掌握标注的内容、标准与流程。在文本标注的整个过程中，项目负责人与标注者始终保持沟通，对标注存在疑惑和歧义的地方及时进行讨论。对于部分缺失的信息，统一用“缺失”两字标注。例如，部分文本的DOI号无法查询，则标注为“<DOI>缺失</DOI>”。标注工作完成之后，三名标注者之间进行了交叉检查，确保标注结果的准确性和正规化。最后，将Word文档批量转换为纯文本文件，并依据总项目要求存放入库。

6. 语料库组织形式

按照总项目要求，文件存放的目录格式为“学科领域\期刊名称\期刊号\论文题目”。其中，学科领域的界定主要参照Web of Science中期刊的分类标签（见表1）。例如，期刊*Aircraft Engineering and Aerospace Technology*归属于航空航天学科（engineering aerospace，编码为HKHTGC）。该期刊2013年第1期题为Gravity gradient torque of spacecraft orbiting asteroids的文章存放目录就是：军事学术英语语料库1.0\HKHTGC\Aircraft Engineering and Aerospace Technology\201301\Gravity gradient torque of spacecraft orbiting asteroids.txt。

需要说明的是，同一种期刊在Web of Science数据库中可能同时被赋予多个领域标签，这种情况下通常以排位第一的学科类别为准。例如，期刊*Terrorism and Political Violence*同时属于国际关系和政治学两个学科，但在本项目中仅选择国际关系作为该期刊的学科领域。

7. 后续建设任务和应用前景

由于该语料库拥有标准的组织结构和规范的建库流程，因此便于进行升级和改造。后续建设任务主要包括两个方面。一是扩大语料库规模。本项目搜集的期刊文献均来自Web of Science数据库，后期将注重收集未被该数据库收录的军事类学术期刊。此外，其他一些军事学术成果的产出形式，如科技报告、学术专著、会议论文等，也可以被纳入语料采集范围之内。二是增加语料加工深度。初步设想是根据Swales（1990）的语步理论对文本中的功能结构进行人工标注。此外，还可以对文本中的汇报动词、模糊语、评价资源、引用点及其功能等方面进行标注。

该语料库的建立具有一系列潜在应用价值。在学术研究方面，该语料库可以为分析军事学术英语的体裁特点提供基础资源。在英语教学方面，该语料库可以为军事英语教材编纂、课堂教学资料设计和学术英语写作平台构建提供一手素材。在词典编纂方面，该语料库可以作为军事学术术语及其例句抽取的重要来源。在自然语言处理方面，该语料库可以为军事学术领域的知识挖掘、信息检索、自动问答系统构建等提供重要支撑。

8. 结语与讨论

语言已经成为一种重要的战略资源，是“自然语言处理战略目标转移的重要标志”（冯志伟 2005）。军事英语语料库的建设与开发是我军信息化建设的必然要求，具有较强的现实性和迫切性。美军尤其注重专用语料库的建设。在美国国防部资助的语音识别、机器翻译、自动语言生成、知识管理等各类自然语言处理研究中，语料库建设都是必不可少的基础性工作。美国国防部一直投入大量的人力、物力构建大规模的标准语料库，为其组织的文本检索（TREC）、信息抽取（MUC）、命名实体识别（MET-2）等会议提供训练和测试语料。为了开展语音识别研究，美国国国防部曾资助构建了King Corpus、TIMIT、ATIS等口语语料库。为了开展机器翻译研究，美国国国防部更是投入了大量资金，重点构建了英语–阿拉伯语，英语–汉语等多种双语平行语料库。与美国等发达国家相比，国内军事英语语料库基础资源的建设与开发还相当薄弱，这在一定程度上影响和制约着我军信息化建设的步伐。

作为“中国外语教育基金专用英语语料库建设项目”的子项目，本课题一方面可以填补国内军事学术英语语料库建设的空白，另一方面该语料库的建库经验可以为后续军事语料库的建设提供参考。更为重要的是，通过参与地方院校组织的高水平语料库建设项目，既有利于提升军事语料库的应用价值，也有利于探索出一条军民融合式的军事语料库建设模式。语料库建设是一个系统工程，军事语

言工作者应着眼“建设信息化军队、打赢信息化战争”目标，加强前瞻规划、顶层设计、统筹联动，为建设自主、可控的高质量军事语言资源做出贡献。

参考文献

Swales, J. 1990. *Genre Analysis: English in Academic and Research Settings* [M]. Cambridge: CUP.

常宝宝、俞士汶，2009，语料库技术及其应用 [J],《外语研究》(5)：43-51。

冯志伟，2005，自然语言处理的学科定位 [J],《解放军外国语学院学报》(3)：1-8。

李永芹、张玉藕、邓全军，2014，军事后勤英语语料库建设的必要性研究 [J],《湖北经济学院学报（人文社会科学版）》(6)：203-204。

王　岚、严灿勋，2015，军事英汉汉英平行语料库建设存在的问题及对策 [J],《解放军外国语学院学报》(5)：33-39。

王　涛、刘文礼、王飞跃、刘　忠，2015，国际军事类学术刊物影响力简评 [J],《指挥与控制学报》(4)：375-383。

肖　健、张音、王宇光、冯占英、张　玉、刘鹏年、赵东升、王松俊，2017，军事医学顶层本体语义关系的构建研究 [J],《军事医学》(11)：929-933。

杨　阳，2011，基于语料库的军事装备术语的翻译研究 [D]。大连海事大学硕士学位论文。

通讯地址：410073 湖南省长沙市 国防科技大学文理学院

中国西班牙语学习者语料库（CACE）：规划与展望*

北京外国语大学 何晓静 刘元祺

提要：各种学习者语料库的纷纷建立，极大地推动了二语习得和外语教学的快速发展。国内外西班牙语学习者语料库的建设实践与相关研究为建设中国西班牙语学习者语料库奠定了坚实的理论、技术和实践基础。在系统梳理当前西班牙语学习者语料库的优势和不足的基础上，本文对中国西班牙语学习者语料库（CACE）的建库理念与基本方案进行了全方位的探讨，并对本库建成后对国内西班牙语的教学与科研带来的影响进行了展望。

关键词：中国西班牙语学习者语料库、学习者语料库、西班牙语教学与研究

1. 引言

自20世纪90年代始，针对外语学习者的各类学习者语料库纷纷建立，并推动了该领域相关研究的蓬勃发展。学习者语料库的构建和研究旨在将基于语料库的数据驱动学习模式融入外语教学和研究。

国内外学界对学习者语料库的研究主要遵循以下三个路径：1）学习者语料库理论与研究综述，如Granger（1996，2003，2004），Mukherjee（2008），李文中、濮建忠（2001），王立非、孙晓坤（2005），邓耀臣（2007），肖忠华、许家金（2008），甄凤超、王华（2010）；2）学习者语料库建库设计和技术研究，如Granger（2012），Alfaifi *et al.*（2014），桂诗春、杨惠中（2003），文秋芳等（2005），卫乃兴等（2007），毛文伟（2009）；3）学习者语言错误和中介语言研究，如Shirato（2007），Guiquin（2011），Schneider & Gilquin（2016），潘鸣威、邹申（2010），何安平、黄雪梅（2011），黄开胜、周新平（2016），潘璠（2016）。从西班牙语学习者语料库的研究来看，Gutiérrez Quintana（2005）、Ramos（2010）、Bailini（2013）、Lozano & Mendikoetxea（2013）、Rojo & Palacios Martínez（2016）等对西班牙语学习者语料库的建库设计进行了介绍；Campillos Llanos（2014）、Lu

* 本文系国家社科一般项目“中国西班牙语学习者语料库的构建与研究”（17BYY122）的阶段性成果。

（2015）、Lozano（2015）等研究聚焦于探讨西班牙语学习者语言输出的某些微观层面；Buyse & González Melón（2013）、Lu & Chu（2013）、Ferraro（2014）、Lu（2015，2016）、García Salido（2016）等则侧重于学习者语料库在西班牙语作为第二外语或外语教学中的应用。

中国的西班牙语专业教学始于1952年，迄今已有65年的历史，全国西语专业的在校学生总人数已经过万。据统计，截至2018年7月，全国共有91所院校设立了西班牙语专业[1]。面对庞大的学生群体，国内西语教学的课程设置、授课方式、教学大纲和语言测等方面均面临紧迫的改革需求。此外，与国内英语及其他语言语言学研究的蓬勃发展相比，国内西班牙语语言学研究的发展仍显滞后。笔者以“西班牙语”为关键词检索词对中国期刊网收录的期刊论文进行检索，发现截止到2018年8月，期刊网仅收录相关期刊论文833篇（其中核心期刊文章仅256篇），且论文选题重复率高，研究方法创新程度不够。排除检索误差和其他因素的影响，期刊网上西班牙语研究论文的低发表率和实证方法的缺席也揭示了当前国内西班牙语语言学研究能力不足的现状。在此背景下，构建一个以中国各大高校西班牙语专业的学生为主体的中国西班牙语学习者语料库（Corpus de Aprendices Chinos de Español，简称CACE）将成为已有西语学习者语料库的重要补充，为全面考察以中文为母语的西语学习者相关指标的研究提供充分的经验数据，为西班牙语二语习得以及中国学生外语习得相关领域的进一步研究提供数据支持，由此推动基于语料库的中国西班牙语学习者中介语的相关研究。

2. 国内外西班牙语学习者语料库现状

根据鲁汶天主教大学英语语言语料库中心制定的《世界学习者语料库一览表》[2]，目前国际上主要的西班牙语学习者语料库共19个，而其中以西班牙语为唯一目标语言的语料库共13个，涉及母语为英语、德语、法语、荷兰语、意大利语等多种语言的西语学习者。语料类型以笔语语料为主，13个西班牙语学习者语料库中，9个为笔语语料库，4个为口语语料库。

笔语学习者语料库以学生作文语料为主，如西班牙圣地亚哥·德·孔波斯特拉大学开发和设计的学习者语料库（Corpus de Aprendices de Español，简称CAES），该库得到了西班牙塞万提斯学院的支持和赞助，采集对象为母语为英语、阿拉伯语、汉语、法语、俄语和葡萄牙语的西语学习者的写作语料。马德里自治大学和格拉纳达大学共同开发了西班牙语二语学习笔语语料库（Corpus Escrito del Español

L2，简称CEDEL2），该库采用在线采集模式，采集对象主要是母语为英语的西语学习者的作文语料，并建立了第二外语为英语的西班牙语本族语者的对照语料库。米兰圣心天主教大学建立了意大利母语者的西班牙语语料库（Corpus del Español de los Italianos，简称CORESPI），该库与西班牙语母语者的意大利语笔语语料库（Corpus del Italiano de los Españoles，简称CORITE）形成对照，两个语料库在数据采集标准、文本数量、学生语言等级以及写作题目等方面完全相同。比利时鲁汶大学开发的西班牙语学习者语料库（Aprendera Escriren Lovaina，简称Aprescrilov）采集了测试和非测试的西班牙语笔语语料，采集对象为鲁汶大学艺术系西班牙语语言和文学专业以及莱休斯应用科学大学应用语言学专业的学生。从西班牙语学习者口语语料库来看，代表性的语料库包括南开普敦大学的西班牙语学习者口语语料库（Spanish Learner Language Oral Corpus，简称SPLLOC）、马德里自治大学语料库小组采集的西班牙语学习者口语语料库（Spanish Learner Oral Corpus，简称SLOC）、西班牙庞培法布拉大学建设的迪亚兹语料库（DÍAZ Corpus，简称DÍAZ）、德克萨斯大学建成的西班牙语熟练水平训练语料库（Spanish Corpus Proficiency Level Training，简称SPT），等等。学习者口语语料库主要采取看图口头表达、命题口头作文、采访、问答等方式进行口语语料的采集。

值得一提的是，中国西语界也在西班牙语学习者语料库建设上取得了较大进展。中国台湾成功大学于2005年开始建设母语为中文的西班牙语学习者语料库（Corpus Escrito de Aprendices Taiwaneses de Español，简称CEATE）。该库从2005年开始建立，最初采集对象全部为专业（第一外语）为英语、西班牙语为第二外语的台湾地区大学生，2008年后语料库增加了4所台湾地区高校西班牙语专业学生的语料。2013年，成功大学西班牙语语料库在笔语语料库的基础上新建了口语语料库（Corpus Oral de Aprendices Taiwaneses de Español，简称COATE），并将笔语和口语语料库合并为台湾地区西班牙语学习者语料库（Corpus de Aprendices Taiwaneses de Español，简称CATE）。

总体来看，国内外现有的西班牙语学习者语料库具有如下特点：

第一，多数西班牙语学习者语料库对于语料采集对象的个人信息（西班牙语水平、国籍、学习地点、所在年级、语言水平、学习时间等）和语料信息进行了标注（句法、词汇、错误类型等）。从采集对象的设置来看，西班牙语学习者包括在校大学生、成年西班牙语学习者以及未成年的西班牙语学习者（如SPLLOC语料库采集了中学生的西班牙语作文语料）。对于西班牙语学习者的语言等级标注，多数语料库选择了《欧盟语言教学与评估框架性共同标准》（Marco común europeo de referencia）和塞万提斯学院制定的语言《课程分级计划》（Plan curricular）作

为语言等级设置标准，CEATE语料库按照学生的学习时间长短进行标注。从语料的加工和处理来看，多数语料库均使用专业软件（Markin、WordSmith、UAM Corpus Tool、CATMA等）用于语料转写、存储、词性标注、句法标注、错误标注、数据检索、结果导出等语料库的构建环节。

第二，语料采集对象的母语以英语或其他印欧语言为主。在《世界学习者语料库一览表》中，将单一亚洲语言设定为调查对象的母语的西班牙学习者语料库仅有两个，分别是CEATE语料库和伯明翰大学的日语母语者的西班牙语语料库（Japanese Learner Corpus of Spanish），其语料采集对象分别为母语是中文和日语的西班牙语学生。CEDEL2、SPLLOC和SPT采集的语料全部来自英语母语者，Aprescrilov的采集对象为荷兰语母语者。CAES和SLOC采集了不同母语背景的西语学习者的语料，但其中仍旧以法语、德语等印欧语言母语者为主，中文母语者的语料很少（CAES中的中文母语者的语料为5万多词，SLOC只采集了四位母语为中文的学习者的语料）。此外，多数学习者语料库都建立了参照语料库，其中，CORESPI和CORITE互为参照库，其基础是母语和目标语的对照（西班牙语与意大利语），而更为常见的是学习者语料库和本族语者语料库构成的参照，如CEDEL2、SPLLOC等语料库，此类库在西班牙语学习者语料库的基础上，也搜集了一定数量的西班牙语本族语者的语料，为学习者和本族语者语料的对比研究成提供素材。

第三，现有的西班牙语学习者语料库整体容量都较小。在上述所列语料库中，Aprescrilov语料库容量最大，为100万词，排名第二的语料库是CEDEL2，其容量为73万词（预计容量为100万词），CAES总容量57万，CEATE语料库的容量为44万词（截至2015年），CORESPI的容量为12万词。鉴于口语语料采集和转写的复杂性，与笔语语料库相比，口语语料库的容量更小，SPLLOC和SLOC的容量都仅为5万词，COATE和DÍAZ口语语料库的容量大小不详。

第四，从上述学习者语料库的开放程度来看，SLOC免费提供语料采集设计、口语任务、话语和错误标记规范、口语语料转写格式等信息，SPLLOC提供所有口语录音和生语料（标记语料不开放）和任务描述，SPT提供视频资料。除此之外，CORESPI、CAES、SPLLOC、CEATE等语料库提供部分语料的免费在线查询，而Aprescrilov等语料库并不对外开放。

综上，一方面，目前国内外已建成的西班牙语学习者语料库的采集对象主要是母语为英语或其他印欧语学习者，未能关注中国西班牙语学习者的主体，因而对于中国西语教学和研究的借鉴意义有限，而语料采集主体为我国大陆地区的西

班牙语专业的学习者语料库建设则尚未展开；另一方面，由于多数研究属于基于自建小型语料库的研究，语料库的原始文本并不对外开放，在线查询和结果输出同样具有一定的限制，故而由语料库触发的相关研究仍旧主要由语料库的开发团队完成，语料库的使用率不高，研究的开放性和普适性有限。

3. 中国西班牙语学习者语料库的建库方案

中国西班牙语学习者语料库（CACE）以中国高校西班牙语专业的学生为语料采集对象，库容约105万词，其中包含学习者语料100万词和本族语者语料5万词。此外，CACE将建立可进行语料检索、提取和加工且具有较强开放性的网络平台。

3.1 语料库构成

中国西班牙语学习者语料库容量约105万词。语料采集的主要对象为中国西班牙语学习者的主体——高校西班牙语专业学生，语料主要来源分为测试作文语料和非测试作文语料两部分，分别构成CACE语料库的测试库和非测试库。测试库的语料来源为全国西班牙语专业八级水平测试的作文语料，非测试语料以命题作文的方式在高校西班牙语专业的学生中采集。此外，参考现有西班牙语学习者语料库的做法，为了最大限度地利用学习者语料库资源，中国西班牙语学习者语料库也构建了参照语料库，采集对象为与中国西班牙语专业学生的教育水平相近的西语国家大学生，并按照学习者语料库和本族语者语料库之间20:1的比例进行语料采集，构建参照语料库。

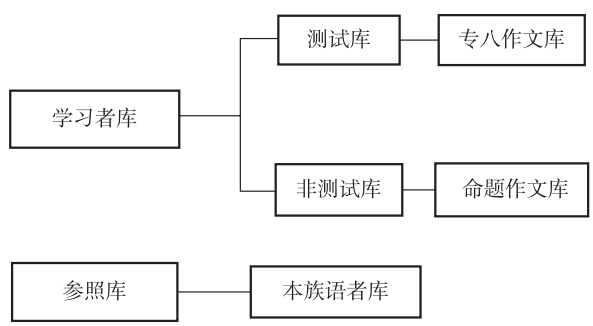


图1 CACE语料库构成

3.2 语料采集

全国高等学校西班牙语专业八级测试于大四上学期举行，是西班牙语专业学生在校期间接受的最高水平专业测试。自2005年全国西班牙语专业八级考试正式

实施以来，考试最后一题均采用命题作文的形式，作文题目对写作的字数进行了规定，一般为200-250字左右。以2014年为分界线，八级测试题目陆续进行了改革。2014年的八级作文并没有规定字数要求。2015年，作文字数限定在150字以上。2016年，八级作文题目进行改版，首次将单一作文分解为应用文和议论文两部分。2018年，应用文的字数规定为50字，议论文的字数为180至220字。整体来看，全国高等学校西班牙语专业八级测试题型相对比较稳定，而除个别年份以外，作文字数要求也比较稳定，因而构成中国西班牙语学习者语料库测试语料库的可靠和稳定来源。课题组按照水平测试考试成绩的等级标准（不及格、及格、良好和优秀）进行分层抽样，建立了中国西班牙语学习者语料库中的测试子库。

虽然测试语料来源稳定，测试作文语料庞大，但是由于考试形式、命题内容、测试对象等都有一定的局限性，因而中国西班牙语学习者语料库也采集了西语专业学生的非测试作文语料，建设非测试作文子库。为了确保语料来源的代表性和均衡性，课题组从全国所有开设西班牙语专业的高等院校中选取六所不同层次的高校进行语料采集。非测试库的作文语料仍旧以命题作文的方式采集，题目主题涉及政治、经济、文化、社会生活等各个方面，体裁包括记叙文、议论文、说明文、应用文和描述文五种。

3.3　语料库标注

完成语料采集和文本清洁后，课题组对语料进行了标注，包括元信息标注、自动词性标注和错误标注三部分构成：

1）元信息标注

作文语料的元信息标注格式如下：

表1　元信息标注示例

元信息	含义说明	示例
<tipo>	语料类型：测试/非测试/本族语者	<tipo>examen</tipo>
<grado>	年级：1/2/3/4	<grado>3</grado>
<fecha>	采集时间：年月日	<fecha>20150510</fecha>
<número>	学号：学校代码+校内编号	<número>011310010</número>
<calificación>	考生得分	<calificación>15</calificación>

2）词性标注

课题组选用广泛应用于西班牙语语词性标注且信度较高的TreeTagger进行自

动词性标注。此外，在自动词性标注完成后，课题组对赋码结果进行了人工修正，以提高词性赋码的准确率。需要修正的错误主要由两类问题引起：其一，软件自带赋码集不全，造成软件无法正确识别（如，前置的间接宾语代词le和les）；其二，因词形相同造成软件无法正确识别词类（如，形容词extraño和动词extrañar陈述式现在时单数第一人称变位词形一样）。诸如以上错误，需在自动赋码后批量修正。

3）错误标注

中国西班牙语学习者语料库课题组对所在院校西班牙语专业高年级作文课学生的作文语料进行了初步分析，并在数轮人工错误标注和一致性检验的基础上进行反复讨论、修正，编制完成较为完善的针对中国西班牙语本科学生作文常见错误的标注体系，编制了相应的赋码集。该赋码集包括12大类、39小类错误，涉及拼写、词汇、句法、词义理解等各个方面。错误赋码集编制完成后，课题组选用北京外国语大学开发的BFSU Qualitative Coder软件进行人工赋码，由此建立了赋码库。此外，对赋码库文本中有2个以上套嵌标注的提取或对整个库的标注内容做批量提取和统计，上述软件可能存在一定困难。本文作者之一，北京外国语大学西葡语系刘元祺博士曾和思科（CISCO）公司技术人员一起开发过一个XML标签内容提取和统计的软件，可以把提取内容和统计结果直接输出到Excel工作簿。因此，在目前阶段，该软件可以较好地解决套嵌标注内容提取问题，但尚有很大改进空间。具体表现在：其一，该软件并不是为中国西班牙语学习者语料库专门开发的检索软件，虽然可以较完美地呈现统计结果，但其统计的部分变量及变量之间的关系可能针对性不强；其二，该软件以命令行方式运行，没有图形界面，操作不直观；其三，功能尚不充实，有待进一步开发。

4. 研究展望

1）语料库综合平台建设

中国西班牙语学习者语料库拟搭建以B/S（浏览器+服务器）架构呈现的网络平台。通过对语料存储、标记、检索和统计分析的统一管理，提高语料库资源的利用率，实现语料库及其软件系统的统一升级。基于网络的检索平台可以大大拓展语料库使用范围，将数据自选存储在服务器，使用者无须安装或下载客户端软件，只需要浏览器就能进行语料的管理和加工，因此具有使用方便、可扩展性强的优点。使用者只要有平板电脑、笔记本电脑、智能手机等能接入因特网的设备，即可上网查询。中国西班牙语学习者语料库拟建成的网络综合平台的结构如下图所示：

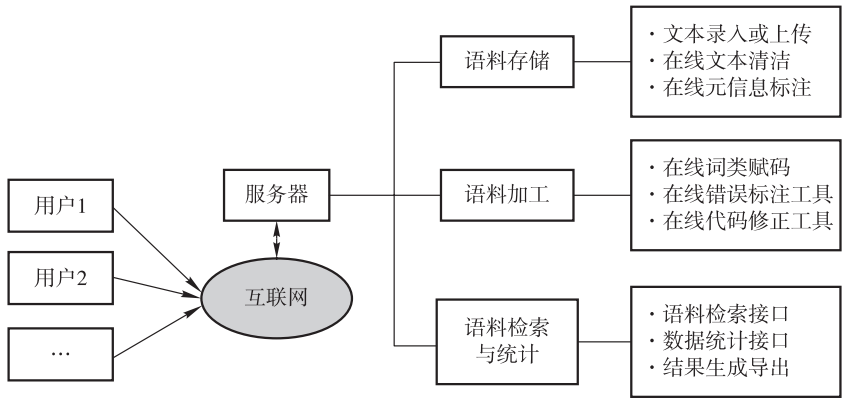


图2　CACE语料库平台结构

2）基于语料库的语言研究

中国西班牙语学习者语料库为从不同维度考察中国西班牙语专业学生的语言习得过程和阶段性特征提供了丰富的语料，并将推动基于语料库的中国西班牙语学习者中介语研究，成为现有的各类西班牙语学习者语料库和大型综合西班牙语语料库（CREA、CORPES XXI、CUMBRE等）的补充，使得不同类型语料库之间的对比研究成为可能。

由于语料库网络平台尚未搭建完成，在现阶段，研究者可以使用本地软件进行检索和结果输出。本库提供UTF-8和Unicode两种编码格式文本，分别支持AntConc和WordSmith Tools这两个常用检索软件。使用者可根据自己实际情况和研究目的，使用其中任何一种在生语料库和赋码库中进行检索。

3）基于语料库的教学研究

基于八级水平测试成绩抽取的作文语料和从各层级院校提取的非测试作文语料保证了语料来源的均衡性，也使得对于不同水平学生和院校的横向对比成为可能。研究者结合语料采集对象所在院校的课程设置、教材使用情况等信息，分析影响和制约中国西班牙语学习者语言能力形成的可能性因素，为西班牙语教学改革、教材和工具书的编写等提供参考。此外，拟搭建的语料库综合平台可以引入西班牙语课堂教学和个人自学，由此带动数据驱动的西语教学。

5. 结论

目前国内外已建成的西班牙语学习者语料库的采集对象主要是母语为英语或其他印欧语言的学习者，未能关注中国西班牙语学习者的主体，因而对于中国西班牙语教学和研究的借鉴意义有限，建设以中文为母语的西班牙语学习者的语料

库已经成为大势所趋。中国西班牙语学习者语料库（CACE）包括高校西班牙语专业学生八级水平测试的写作测试语料、以命题作文形式所采集的非测试语料以及从本族本族语者采集的笔语语料三部分，并将建成一个具有较大容量、较强开放性的语料库平台。研究者可以根据不同的教学和科研需求，展开基于语料库的相关课题的研究，将有助于推动研究和教学法的创新，并进一步服务于我国的西班牙语教学与科研。

注　释

1.《第五届全国高校西班牙语教学观摩研讨会在京举行》，CRI国际在线，2018年7月19日，http://news.cri.cn/20180719/4ba412ec-e627-13ae-9eab-d62ad207672e.html.

2. Louvain-la-Neuve: Université catholique de Louvain, Centre for English Corpus Linguistics: “Learner Corpora around the World.”

参考文献

Alfaifi, A., A. Eric & I. Hedaya. 2014. Arabic Learner Corpus (ALC) v2: A new written and spoken corpus of Arabic learners [A]. In S. Ishikawa (ed.). *Learner Corpus Studies in Asia and the World* [C]. 77-89.

Bailini, S. 2013. SCIL: A Spanish corpus of Italian learners [J]. *Procedia-Social and Behavioral Sciences* 95(4): 542-549.

Buyse, K. & E. Melón. 2013. El corpus de aprendices Aprescrilov y su utilidad para la didáctica de ELE en la Bélgica multilingüe [A]. In S. Borrell, B. Falgaeras, B. Crous & F. Sierra (eds.). *Plurilingüismo y enseñanza de ELE en contextos multiculturales: XXIII Congreso Internacional ASELE* [C]. 247-252.

Campillos Llanos, Leonardo. 2014. Análisis de errores pragmático-discursivos en un corpus oral de español como lengua extranjera [J]. *Círculo de Lingüística Aplicada a la Comunicación* 58 (2): 23-59.

Ferraro, G., R. Nazar, M. Ramos & L. Wanner. 2014. Towards advanced collocation error correction in Spanish learner corpora [J]. *Language Resources and Evaluation* 48(1): 45-64.

García Salido, Marcos. 2016. Error analysis of support verb constructions in written Spanish learner corpora [J]. *The Modern Language Journal* 100(1): 362-376.

Gilquin, G. 2011. From EFL to ESL: Evidence from the International Corpus of Learner English [A]. In J. Mukherjee & M. Hundt (eds.). *Exploring Second-Language Varieties of English and Learner Englishes: Bridging a Paradigm Gap* [C]. Amsterdam: John Benjamins. 55-78.

González, A. & M. Ramos. 2013. A comparative study of collocations in a native corpus and a learner corpus of spanish [J]. *Procedia-Social and Behavioral Sciences* 95(4): 563-570.

Granger, S. 2003. Error-tagged learner corpora and CALL: A promising synergy [J]. *Calico Journal* 20(3):465-80.

Granger, S. 2004. Computer learner corpus research: Current status and future prospects [J]. *Language & Computers* (1): 123-145.

Granger, S. 2012. The international corpus of learner English: A new resource for foreign language learning and teaching and second language acquisition research [J]. *TESOL Quarterly* 37(3): 538-546.

Granger, S., G. Gilquin & F. Meunier 2013. *Twenty Years of Learner Corpus Research: Looking Back, Moving Ahead* [M]. Louvain: Presses Universitaires De Louvain.

Granger, S., G. Gilquin & F. Meunier. 1996. *From CA to CIA and Back: An Integrated Approach to Computerized Bilingual and Learner Corpora* [M]. Lund: Lund University Press.

Gutiérrez Quintana, E. 2005. Corpus de textos escritos por universitarios italianos estudiantes de ELE [J]. *Lingüística en la Red* (3):1-52.

Mukherjee, J. 2008. English corpus linguistics and foreign language research: Line of development and perspectives [J]. *ZFF, Zeitschrift fur Fremdsprachenforschung* 19(1): 31-60.

Lozano, C. 2015. Leaner corpora as a research tool for the investigation of lexical competence in L2 Spanish [J]. *Journal of Spanish Language Teaching* 2(2):180-193.

Lozano, C. & A. Mendikoetxea. 2013. Learner corpora and Second Language Acquisition: The design and collection of CEDEL2 [A]. In A. Díaz-Negrillo, N. Ballier & P. Thompson (eds.). *Automatic Treatment and Analysis of Learner Corpus Data* [C]. Amsterdam: John Benjamins. 65-100.

Lu, H. 2015. Estudio de verbos copulativos a partir de corpus de aprendices [J]. *Porta Linguarum* 23(1): 205-220.

Lu, H., S. Hung & L. Lu. 2016. La aplicación de un corpus de aprendices en la autocorrección de composiciones escrita [J]. *Porta Linguarum* 26(1): 149-160.

Lu, H. & Y. Chu. 2013. Evaluation of corpus-assisted Spanish learning [A]. In *Proceedings of the 27th Pacific Asia Conference on Language, Information and Computation (PACLIC 27)* [C]. 467-473.

Ramos, M., L. Wanner, O. Vincze, G. del Bosque, N. Veiga, E. Suárez & S. González. 2010. Towards a motivated annotation schema of collocation errors in learner corpora [A]. In *LREC 2010 Proceedings* [C]. 3209-3214.

Rojo, G. & P. Martínez. 2016. Learner Spanish on computer. The CAES "Corpus de Aprendices de Español" project [A]. In M. Ramos (ed.). *Spanish Learner Corpus Research: Current Trends and Future Perspectives* [C]. 55-87.

Shirato, J. & P. Stapleton. 2007. Comparing English vocabulary in a spoken learner corpus with a native speaker corpus: Pedagogical implications arising from an empirical study in Japan [J]. *Language Teaching Research* 11(4): 393-412.

Schneider, G. & G. Gilquin. 2016. Detecting innovations in a parsed corpus of learner English. International [J]. *International Journal of Learner Corpus Research* 2(2):177-204.

邓耀臣，2007，学习者语料库与第二语言习得研究述评 [J],《外语界》(1)：16-21。

桂诗春、杨惠中，2003,《中国学习者英语语料库》[M]。上海：上海外语教育出版社。

何安平、黄雪梅，2011，英语教材话语的立场标记语探究 [J],《当代外语研究》(3)：10-16。

黄开胜、周新平，2016，基于语料库的中国英语学习者词块输出能力的趋势研究 [J],《外

语界》（4）：27-34。

李文中、濮建忠，2001，语料库索引在外语教学中的应用 [J],《解放军外国语学院学报》（2）：20-25。

毛文伟，2009，整合型学习者语料库平台的规划与实现 [J],《现代教育技术》（9）：54-61。

潘鸣威、邹　申，2010，英语专业高年级学习者写作中读者/作者显现度再探 [J],《外语教学》（4）：48-52。

潘　璠，2016，语料库驱动的英语本族语和中国作者期刊论文词块结构和功能对比研究 [J],《外语与外语教学》（4）：115-123。

王立非、孙晓坤，2005，国内外英语学习者语料库的发展：现状与方法 [J],《外语电化教学》（5）：19-24。

卫乃兴、李文中、濮建忠，2007，COLSEC语料库的设计原则与标注方法 [J],《当代语言学》（3）：235-246。

文秋芳、王立非、梁茂成，2005,《中国学生英语口笔语语料库（1.0版）》[M]。北京：外语教学与研究出版社。

肖忠华、许家金，2008，语料库与语言教育 [J],《中国外语教育》（5）：51-52。

甄凤超、王　华，2010，学习者语料库数据在外语教学中的应用：思想和方法 [J],《外语界》（6）：72-77。

通讯地址： 100089 北京市 北京外国语大学西葡语系

《语料库语言学在翻译与对比研究中的应用——研究指南》述评

北京外国语大学 詹潇潇

Mikhail Mikhailov & Robert Cooper 2016. *Corpus Linguistics for Translation and Contrastive Studies: A Guide for Research.* London: Routledge. xxiii+233pp.

1. 引言

通过语料库揭示语言使用规律已渐成主流。语料库也被用于分析翻译语言，这标志着翻译学从传统的规定性研究走向描写性实证研究的转变。一直以来，单语种和类比语料库备受语言和翻译研究者的青睐，而平行语料库由于文本获取和对齐难度较大，发展相对滞后。在此背景下，该书详细介绍语料库方法在翻译与语言对比研究中的应用。

2. 内容简介

该书共有七章。第一章简要介绍了语料库的类别、存在的问题和语料库的作用。双语语料库在翻译领域用途广泛，既可协助译者的工作，也可用于研究翻译过程、翻译策略、译者风格，在机器翻译技术层面也大有作为。尽管如此，双语语料库的发展滞后于单语种语料库，原因在于与原文对应的多语种译文难以获得，且在对齐上难度较大，费时费力。此外，语料库本身也存在缺点和问题，比如有些语法形式缺失或出现频数过低。另外，习语、隐喻、讽刺、幽默等特殊表达只能依靠手动标注。

第二章涉及如何设计和创建平行语料库。建库者首先要准备好所需的软件和硬件设备。创建平行语料库主要有以下步骤：1.设定好语料库的规格；2.列好所需的文本类型；3.输入电子文本；4. 对齐；5.标注；6. 存储；7. 版权问题（pp. 53-54）。建库之前，建库者需要确定好语料库的规模以及元信息（metadata）的具体构成（包括文本类型、体裁、领域、创作时间和作者等）。其次，要获得电子文

本，可以手动输入，也可扫描纸质版后进行文字识别。下一步为对齐文本，这是建立双语语料库的重要步骤，但手工对齐费时费力，可使用Trados中的WinAlign、Déjà Vu、bitext2tmx等文本对齐工具自动对齐，之后进行人工校对。

文本标注分为文本外部信息标注和内部信息标注，前者即所谓的元信息，后者指的是文本切分、文本内部关系和语用信息等。文本的标注层次和类型因建库目的不同而各异。最后一步是存储建好的语料库，可存成本地文档，也可做成在线形式，如英国国家语料库（BNC）就采用了这两种形式存储。此外，建库者还要考虑搜集的文本是否涉及版权问题，需要获得文本创作者的许可，尊重知识产权，建好的语料库的传播和使用也同样如此。

第三章介绍使用语料库的基本步骤。在语料库中检索需要借助语料库检索语言（Corpus Query Language）。如BNC语料库使用的就是CQP（Corpus Query Processor）检索语言,需要用到正则表达式（regular expression）。正则表达式是用形式化语言编写的表达式，用于匹配符合某个句法规则的字符串（比如某个字、词或型式），在语料检索中发挥着重大作用。

语料库检索结果通常是以索引行（concordances）的形式呈现。索引行是指一串词或短语共同出现在即时语境中（pp. 93），通常以语境中的关键词（key word in context，KWIC）形式呈现。语料库另一个主要用途是生成词频表，这是多数基于语料库研究的开端。Leech（2011：8）认为频数是语料库提供的最有用的信息。搭配词是指经常在同一语境下出现的词组，如keep + talking、refuse + to talk。搭配词有诸多用途，比词频和索引行更能说明词的用法和语义，可用来解释语言变体之间的不同。

第四章从具体的个案出发，探讨如何处理语料库检索结果：如通过平行索引行比较翻译对等项、生成词频表、研究搭配词在单一语言或平行语料库中的使用。虽然平行语料库设计的初衷是用于跨语言研究，但对单一语言的研究用处也很大。本章列举的探究英文中地点介词before和in front of的用法，就用到了文学语料库Farkas和Linguee双语网站（广义的语料库包含网站）。检索结果显示，从19世纪开始，地点介词before逐渐让位于in front of，但在司法和宗教文本中before出现的频次很多，并常出现在appear等动词后。

第五章分析了语料库中的统计方法，介绍了两种常用统计软件：SPSS和R，论述了测量数据可靠性的必要性，引入了测量集中趋势和离散趋势的统计方法。测量数据的可靠性（reliability test）是基于语料库研究的第一步。在创建平行语料库时，由于各个作家作品的字数不同，所以很难构建平衡的全文语料库。作者建议可根据需要，建立多个子语料库，使数据更接近自然分布，研究结果也更可靠。所以在创建语料库时，不仅要考量库容，还需注意其构成和代表性。在介绍了平均数、中位数、众数、最大/最小值、标准差、峰度、偏斜度、方差和Z值等测

量数据分布的参数后，作者介绍了标准类符形符比（standardized type/token ratio，STTR）、平均句长、TamBiC中常见词的分布等参数的计算方法。作者发现，不同译者译文的STTR与译者风格的相关度高于与原文风格的相关度，这印证了Baker（2000：249-250）提出的STTR主要由译者风格所决定的观点。但作者同时指出，还需通过研究不同语言对的译本数据来证实或证伪这一观点。

第六章探讨平行语料库在词典学、术语学、构词学、句法学、语用学、翻译学等领域的应用。平行语料库可用作网上词典，也可存为翻译记忆库,为译者翻译或外语写作提供便利。作者也指出多语种语料库未来的研究方向，如跨语言研究、翻译评估、性别语言研究等。从语料库中提取的词频表、词语使用的具体语境和搭配词，可用于词典编纂。但也存在一些问题，如并非所有文本类型的双语文本都能得到，这导致大型平行语料库中文本代表性不足；其次，建库者获取的数据并非完全可靠，译本的质量也良莠不齐。因此在编纂双语词典时，完全依赖平行语料库是不够的，还需参照单语语料库获取更多数据。

语料库在翻译研究中用处也很广，可用于译者培训和机器翻译的准确性检测。由译本和译入语原创文本构成的单语种语料库可用于研究翻译语言与原创语言的差异，一个很好的例子就是由曼彻斯特大学创建的英文翻译语料库（Translated English Corpus）。而平行语料库则更多地用于对比研究。Olohan（2004：13-14）指出基于语料库的这类研究忽略了翻译的实际过程，没有考虑到翻译的目的、译者的背景、受众等，只着眼于语言和风格，所以要在语言对比和翻译研究中寻求宏观外部信息（contextual）与微观文本层面（textual）的平衡，并增加平行语料库在翻译过程研究中的使用。

第七章概述现有的平行语料库。多语种平行语料库主要有欧洲的各种对齐语料库、法律文件多语种语料库（a multilingual corpus of legal documents）以及亚洲语言与英语语言对的平行语料库，包括北外中英平行语料库（BFSU Chinese-English Parallel Corpus）、巴别塔中英平行语料库（the Babel Chinese-English Parallel Corpus）等双语平行语料库。口译语料库在跨文化交际中也有所应用，这类语料库主要有：Bologna大学创建的欧洲国会口译语料库（The European Parliament Interpretation Corpus)、交传和同传语料库CoSi、Dolmetschen im Krankenhaus医院口译语料库等。

3. 述评

该书有以下特点：

第一，实用性强。正如此书的标题所言，该书是介绍如何利用语料库语言学进行翻译与对比研究的实用指南，展示了平行语料库和多语种语料库的创建步骤，

介绍其在词典编纂、语言比较以及译者培训与翻译研究中的应用。作者首先介绍了建库的主要步骤和注意事项，接着以丰富的案例展示如何利用各种类型的语料库进行检索，并借助统计学知识测量数据、解读数据，保证数据的可靠性。每一章节都有导论，表明该章节论述的话题和目的，解释该章节的一些核心概念，同时和前后章节都有所呼应；每一章节在最后都总结了该章节的主要内容，并列有注释和参考文献，方便读者理解和进一步查阅相关文献。第六章重点探讨语料库在词典学、术语学和翻译领域的应用，除了给出具体的案例，还针对每个领域展望基于语料库的研究课题和方向，激发读者的思考和研究兴致。

第二，案例丰富翔实。无论是DIY语料库的设计和创建，还是对检索结果的阐释和独立性、可靠性检验，作者都列举并分析了芬兰语–英语、芬兰语–俄语等不同语言对的案例来说明上述问题。个案全面详尽，步骤清晰，可操作性和可复制性强，为读者研究其他语言对提供了范例，研究视角和方法也值得借鉴。

第三，体现跨学科视角。本书把语料库语言学和翻译与语言对比研究相结合，并借助计算机编程技术和统计学来攻克语言对比和翻译研究的难题，体现了翻译学和语言对比研究的跨学科、多维度的范式和本质，足见未来各学科的渗透、融合的发展态势。

但本书也存在一些不足之处。首先，受作者自身学术和语言背景的限制，文中所举的案例大多围绕芬兰语–英语、芬兰语–俄语等语言对展开，这给不熟悉这些语言的读者增加了信息处理负担。作者在文末也承认此不足，并解释道：个案的目的并非向读者介绍某一语言的新知识，而是向他们展示如何使用一些研究方法和视角去研究任何一组语言对，所以并不要求读者有丰富的芬兰语或俄语知识。其次,就像作者在结语中提到的，第七章列举的现存的平行和多语语料库都是欧洲语言对之间的，亚洲和非洲语言的语料库鲜有提及，作者还指出后者的建设还处在基础阶段。但就中文而言，据笔者所知，中文的单语和平行语料库，包括书面语、口语形式，其实不在少数。如20世纪90年代清华大学计算机系的“现代汉语语料库”，总库量超过1000万词，以及中文系建立的“ZW大型通用汉语语料库”；1992年北京语言大学建成“当代北京口语语料库”、中文语言资源联盟（Chinese Linguistic Data Consortium）、国际中文语言资源联盟（Chinese Corpus Consortium）；武汉大学建立的汉语现代文学作品语料库（1979年，527万字）；北京航空航天大学的现代汉语语料库（1983年，2000万字）；北京师范大学的中学语文教材语料库（1983年，106万8千字）；香港城市大学语言资讯科学研究中心的LIVAC（Linguistic Variation in Chinese Communities）语料库。双语语料库有英汉双语语料库、日汉对译语料库、德汉双语语料库、汉日英分类熟语料库等。与此同时，还建立了包括维吾尔语、藏语、蒙古语等少数民族语言语料库。可见汉语的语料库发展蔚为壮观，而原书作者偏重印欧语系的语料库研究，而对包括汉

语在内的亚洲语言、非洲语言的语料库的建设关注度不够、掌握的资料也不够全面充分，故而观点也有失偏颇。

4. 结语

总而言之，该书瑕不掩瑜，内容丰富、案例翔实，是一本非常实用的语料库语言学研究入门专著，对翻译和语言对比研究者都有较高的参考价值。很多人觉得操作各种计算机程序软件很是复杂、难以上手，而该书基本可以回答关于语料库概念、建设、检索、应用等方面大部分问题，通过丰富有趣的个案讲解，使抽象的概念和繁杂的操作步骤变得具体、清晰，值得一读。

笔者认同作者在文末提到的观点，即：整个世界越来越数字化，电子文本的获取、标注和处理变得越来越简单、快捷，未来任何对语言研究感兴趣的人创建语料库将成为一种惯例（p. 303）。在互联网和计算机技术日渐发达的今天，要想从海量、繁芜复杂的数据中高效便捷地辨识和提取有用信息，掌握语料库检索技术以及统计学相关知识很有必要。

参考文献

Baker, M. 2000. Towards a methodology for investigating the style of a literary translator [J]. *Target* 12(2): 241-266.

Leech, G. 2011. Frequency, corpora and language learning [A]. In F. Meunier, S. de Cock & G. Gilquin (eds.). *Taste for Corpora: In Honour of Sylviane Granger* [C]. Amsterdam: John Benjamins.

Olohan, M. 2004. *Introducing Corpora in Translation Studies* [M]. London: Routledge.

通讯地址： 100089 北京市北京外国语大学英语学院

English abstracts

The discursive construction of Chinese stock market image by the Western media *as per* the attitude system

Drawing on the attitude system of the Appraisal Theory and using corpus-based Critical Discourse Analysis, this article explores the image of Chinese stock market constructed in the mainstream Western media. The research is based on the news reports of Chinese stock market covered by the mainstream Western media in the recent five years. The results show that while acknowledging the achievements of Chinese stock market, the Western media strongly criticized the immorality of China's economy. In addition, the Western media believe that Chinese investors are blind, and rebuke the management of Chinese stock market and the government's administrative intervention measures regardless of the development stage of Chinese stock market. This discourse model reflects the influence of Western ideology on its media, as well as its constant propaganda of its "free market" system.

A corpus-based contrastive study of Chinese adverbs *zhǐ* (只) and *guāng* (光)

Based on the Chinese National Corpus, this study analyzes the similarities and differences between Chinese synonymous adverbs *zhǐ* (只) and *guāng* (光) in terms of their syntax, semantic functions and pragmatic entailments. The results show that whether *zhǐ* (只) and *guāng* (光) can alternate is influenced by such factors as their semantic prosody, monotone entailment, semantic suitability and subjective small quantity.

Eileen Chang's self-translation of *Yuannv*: Its lexical features and publication plight

The end product of literary translation is its publication. Despite that publication is by no means the only judgment of translating activities, the content of translated works is very likely to have an impact on it. In the case of Eileen Chang's self-translation of *Yuannv*, two English versions have been composed, yet only *The Rouge of the North* got published, leaving *Pink Tears* only historically recorded yet unpublished. Previous studies contradict with each other in describing the sequential order of *Yuannv*'s bilingual texts, which triggers the present investigation into its

translation direction and hence a corpus-based study of the lexical features of *The Rouge of the North* and an analysis as to why *Pink Tears* was denied the publication opportunity based on Eileen Chang's English writing and publishing principles.

Inconsistent tense use in reported speech in translated English news

In view of the importance of tense use in news discourse, the research focuses on inconsistent tense use in reported speech of Chinese-English translation of newswire texts. We examine the overall tense features and semantic patterns of past tense reporting verb and intentional absolute present reported verbs. Furthermore, the source text effects are explored from syntactic, semantic and pragmatic levels to achieve a systematic description and analysis of the reasons behind inconsistent tense overuse in translational English from Chinese. It is found that the inconsistent tense collocations are significantly more than in CNN native English news. Such overuse is due to "source text effects" from syntactic, semantic and pragmatic levels.

A study of Chinese EFL learners' authorial identity prominence in academic English writing

Authorial identity helps to facilitate discourse communication in academic writing, reporting clauses are an important resource for highlighting authorial identity. By comparing to the corpora of international journal articles, this paper focuses on the subject of reporting clauses and explores the authorial identity features of Chinese EFL learners in academic writing and their developmental characteristics in undergraduate, master and doctorate stages. Results show that the learners' awareness in showing authorial identity is weak. As for the developmental characteristics, such identity is on the rise as the learners' academic level goes up on the whole. But there is great difference in specific usage. For example, non-standard uses of reporting clauses are found in undergraduate students' texts. Master students tend to cite others' research to add to their discoursal credibility. Doctorate students show their relatively mature authorial identity in using reporting clauses.