

Three-dimensional Model Construction of Sports English Referencing on Corpus Linguistic

Shi Rong

Guangdong University of Science and Technology, Dongguan, Guangdong, 523000, China

Abstract

As an emerging discipline, Sports English has attracted more and more scholars' attention. With the method of corpus linguistics in widely used in the study of English for special purposes, the researchers began to explore the construction of sports English corpus problems, such as workers Democratic Party Chen jiansheng, qing-song tan discusses the problem of the relationship between the sports English corpus and ESP, through comparing with general English, on the sports English linguistic orientation. Although this study emphasizes the necessity of classifying the corpus from the perspective of linguistic features, it fails to further discuss the principles and methods of text selection in sports English corpus. At the same time, sports corpus began to appear, such as MUFEN corpus focusing on Manchester United football news. When researching thematic Corpus, people often need a Reference Corpus, or comprehensive Corpus for comparison. Baker et al. describe a referential corpus as "a larger set of texts collected from more registers and/or more sources". However, "balance, representativeness and comparability are the ideal that all corpus builders try to achieve, but it is difficult to achieve". Few researchers have proposed solutions on how to construct a reference corpus. This study tries to put forward an operational framework for building sports English reference corpus from three dimensions: the social attributes of sports discourse, the register characteristics of sports language and the quantitative principle of corpus selection.

Keywords

Sport English; Corpus Linguistic; Three-dimensional model.

1. Social Dimensions

The sociality of sports reflects the attribute of sports as human social experience. Lu Feng [5] pointed out, "Sports is a social and cultural form created by human beings according to their own needs. It exists in the cultural mode of each specific era or nation, and adjusts its content and structure and transforms its functions with the changes of human society's needs. Sport is created in history, co-exists among nations and is inherited in cultures, with various forms and contents. Sport is also linguistic, reflected, recorded and transmitted by language. The discourse value reflected by sports language is the core content of sports social attribute. With the help of a large number of corpus facts, people can not only review the development of sports, but also compare the different group consciousness reflected by national sports, and observe the interaction between sports and economy. The following will analyze the social attributes of sports in detail, and classify sports events in the form of charts, so as to clarify the main sports events and clarify the objects of corpus sampling.

1.1. Historic History of Sports

Physical activities originated from human demands for life, and sublimated them through art forms. For example, running comes from people's pursuit of speed, throwing comes from human hunting experience, windsurfing reflects people's tribute to their ancestors' sailing skills and

so on. With the development of society, the form and content of sports have also become more and more rich. In a broad sense, mental items unrelated to physical skills and dexterity, such as chess and games, are also included in sports. Sports participants are no longer only a few people engaged in competition. With the expanding scale of urbanization, more and more citizens for fitness purposes actively participate in sports. The change of sports constitutes an important part of the development history of human society.

1.2. Ethnic Nature of Sports

Farmers are familiar with utensils, herdsman are good at riding, and fishermen are skilled in boats. Sports was born in the process of human production, and is closely related to the mode of production. The original form of sports is directly derived from human production activities, and has a strong regional and ethnic nature, forming the ethnic difference of modern sports. On the one hand, people in different countries and regions may not have the same sports hobbies, such as Americans are keen on football, while the British flock to cricket and football. On the other hand, with the process of political and economic globalization, many sports also gradually move from regionalization to globalization, such as the Olympic Games are committed to emphasize the globalization of sports and weaken their ethnic nature. Some sports, such as football, have become a global sport, such as the UEFA Champions League, the Asian Champions League, the South American Cup and other professional events have become an important part of the regional sports discourse. The quadrennial World Cup has been popular around the world for dozens of years. However, the regional and global differences in sports remain.

1.3. Cultural Nature of Sports

The culture of sports reflects the connotation and characteristics of sports, mainly manifested as competitive culture and leisure culture. Stefani [6] believes that sports can be divided into three categories: control opponent type (Combat Sports), control target type (Object Sports), and independent type (Independent Sports). Control of opponent sports refers to events that surpass the opponent in speed, height, distance with or without equipment, such as track and field, rowing, etc. Control target sports refers to items in which competitors compete for each other for the same goal to score, mainly ball sports. Independent sports refers to sports that do not target sports, including a variety of leisure self-help fitness puzzle sports. Obviously, the first two belong to the competitive culture, "confrontation" is its core discourse characteristic [7], the latter belongs to the leisure sports culture.

1.4. Economy of Sports

"Sports has always been linked to business, but there has been an unprecedented increase in the number, frequency and intensity between the links over the last 25 and 30 years" [8]. The economic performance of sports is the distinction between "professional" and "non-professional" sports, among which "non-professional" is divided into "competition" and "non-competitive". The development of professionalism and competitive sports has led to the construction and operation of sports venues, the production and sales of sporting goods, as well as the development of related services (such as advertising, gambling, etc.). Large-scale sports events, such as major professional leagues and the Olympic Games, have promoted the social and economic development and promoted urban construction. According to the social attributes of sports, all sports can be positioned through a social-dimensional network system (see Figure 1). Sports can be roughly divided into seven categories, each with several items. For example, the social characteristics shared by football and basketball are: global + control target type, equal confrontation + professionalism or competition type.

2. Language Dimension

The language dimension mainly considers two questions: where does the corpus come from? How to choose the corpus?

Although English has become a cosmopolitan language, the frequency and proficiency of varies in different countries and regions due to reasons such as history and culture. To distinguish between different English communities, Kachuru[9] proposed "English Third Ring Model (Three Circles of English): the inner ring includes several native English speakers: Britain, USA, Canada, Australia, and New Zealand; Central includes a bilingual society with English as one of its official languages, mainly in British former colonies, such as Singapore, India, Hong Kong, etc.; all countries that regard English as foreign languages, such as China, Japan, Brazil, etc. The reference corpus is not subject to specific research purposes, and its nature determines that the corpus must be sourced only to native English-speaking countries, the inner ring in the third ring.

Nationality Character	Globality		Category 1: Rowing, Racing, horse racing, cycling, throwing, weight lifting, fencing, shot put, javelin, discus, shooting, archery, skating, snow, gymnastics, etc
	Regional		Category 2: professional leagues such as NBA, NFL, and various regional professional leagues.
Cultural Character	Controls the opponent type	With equipment	
		With Non-equipment	Category 3: marathon, long-distance running, sprint, race walking, high jump, long jump, boxing, wrestling, judo, taekwondo, swimming, diving, etc.
		Stand-alone type	
	Control the target type	Net Dividing Contact	Category 4: professional leagues, such as the Australian Open, Open, French Open, US Open, etc.
		Same-Court Contact	Category 5: basketball, football, football, ice hockey, training ball baseball handball,, softball, golf, etc.
Economical Character	Professionalism		
	Non-professional	Competition type	Category 6: table tennis, volleyball, tennis, badminton, etc.
		Non-competition	Category 7: leisure sports such as jogging, hiking mountain climbing, rock climbing, limit.

Figure 1. Classification of sports items based on sports society

Corpus acquisition adheres to domain principles (Register Principle) [10-12]. Language domain refers to the diversity of language communication functions, is "a meaning configuration that is closely related to a certain context of language field (Field), language style (Mode), Fundamentals (Tenor) configuration" [13]. The language domain realizes context at the

semantic layer, and also specifically at the lexical syntactic layer. Matthiessen et al. [14] provides a very intuitive criterion for discourse type discrimination in the classification of social meaningful processes based on language fields. In this mode, the social meaning process is divided into eight domains, including interpretation, reporting, and reproduction, each of which is embodied as two different semantic features by dialogue and monologue, and each semantic feature is realized by different discourse types. This uses this story type mode to locate sports text types (see Table 1).

As shown in Table 1, some language domains are very limited in sports discourse, such as the "Representation: Narrative" category. This is because this text is often literary works like fiction, poetry, and opera, while fictional sports genres are relatively rare. This phenomenon is called "domain deletion (Missing Register)", "the missing domain is because in the language community, the corresponding language scenario does not appear, precisely, the combination of themes and activities, relationships and discourse patterns do not appear. This reflects and guarantees the stability of the social system of the language community" [15]. In a sense, the phenomenon of domain loss is one of the important features of specialized English different from general English.

When collecting the corpus, attention is also necessary to distinguish the "closed" from the "open" domains. Some domains need to appear in very specific contexts, such as the referee must appear at the game scene, and are highly stylized, short and not ambiguous. This type of domains is known as the "closed language domain" [15]. Closed domains, due to their limited number, can be exhausted during corpus collection. However, when collecting a corpus, it still lacks any quantitative advantages over all the relevant ones, even if exhausted. However, from the perspective of research or learning, the text of the closed language domain is relatively easy to grasp because of its high degree of programming. Therefore, adopting a corpus approach for studies of such domains is not desirable. In contrast, open domains have less controllability as their variables increase. This means that a large amount of corpus is needed to cover the domain. For example, football news involves from the players to the game, from the season to the playoffs, from the field to outside the field, involving a wide range of areas, far more than the closed language area. The general requirement is that the closed the domain, the less the corpus; the more open the domain, the more the corpus is needed to ensure the equalization of the corpus.

Combined with the classification of sports from the social dimension in the previous section, on the basis of the built comprehensive sports English corpus, the corpus users can build both from the research and teaching needs from the type of discourse. For example, if you build a football English corpus from sports, first select football as the default item in the Figure 1 system, and then clarify the type of discourse contained in the football English corpus in Table 1, and determine the architecture of the subcorpus (Figure 2). If the English corpus of sports news is constructed from the discourse type, first select the sports news as the silent option in Table 1, and then select the relevant sub-libraries of the sports items listed in Figure 1 system one by one to determine the structure of the corpus (Figure 3). For example, "NBA event report corpus [7]" combines the social characteristics of sports and sports, and the inclusion conditions of the corpus search scope are: "Sports News: Written: Regional (USA): Professional: control: opponent type: (same match): basketball".

Table 1. Types of Texts for Language Areas and Sports English

register	Semantic features	Language	
		the types of discourse; text type	
		written language	Oral language
(expounding)	(explaining)	PE knowledge	knowledge lecture
	(categorizing)	Physical education teaching materials	
(reporting)	(surveying)	census	interview
	(chronicling)	News report	PE Broadcast
(recreating)	(dramatizing)		Sports theme film and animation
	(narrating)	Story	
(sharing)	(sharing experiences)	Biography	
	(sharing value)	Event Q & A blog	
(doing)	(collaborating)	Event arrangement	
	(directing)	Coach guidance	The referee instructing & the coach guiding
(recommending)	(advising)	sports advertisement	
	(promoting)	Event publicity	television advertising
(enabling)	(instructing)	Sports rules	
	(regulating)	Sports policy rules and regulations	
(exploring)	(debating)	Sports papers	On-site commentary, the academic forum
	(reviewing)	Event comments	Speech

Football English corpus						
written language				Oral language		
Academic Paper	Football teaching materials	Football news		Academic forum	On-site commentary	Interview

Figure 2. Football English corpus-sub-library

Sports News English corpus						
written language				Oral language		
Football	Basketball	Swimming		Football	Basketball	Swimming

Figure 3. Sports News English corpus-Sub-library

3. Quantification Dimension

Although the equilibrium, representation, and comparability of corpora are difficult to achieve, they are still essential elements of the quantitative dimensions of the corpus. In solving the equalization of the corpus, the corpus builders mainly applied the "quantity averaging principle", such as the Brown corpus. The Brown corpus has 500 sampled text, with 2,000 words each, totaling one million words. For a corpus built for a particular purpose without the need to consider the diversity of language types. Another principle is the "language principle", such as the British National corpus (BNC), which classifies both verbal and discourse types. The BNC has been compared to both a "garden" and a "jungle" [16], stating very vividly that the corpus classification is not an easy thing. Obviously, it is difficult for building a comprehensive corpus to meet the principle of corpus representation by using any classification method alone (see 3.2). New schemes are needed to bring the composition of the integrated sports English corpus closer to representation and equilibrium.

3.1. Equilibrium of the Corpus: The Weight Ratio of Sports Discourse

Sports discourse is mainly created, transmitted and recorded through modern media, education, scientific research and other media and channels. The essence of the discourse is its sociality. In terms of the sociality of a whole, economic factors (or the commercialization of modern sports) play a decisive role in the promotion and development of modern sports, so it also determines the weight ratio of sports discourse. From an economic point of view, the weight ratio of sports discourse is successively professional sports, competition sports and non-competitive sports. Professional competitions tend to invest a lot of money in the publicity and operation of events. From the media coverage object, the coverage of professional events is the most lasting, the most comprehensive. Second, the attention to non-professional large sports events (such as the Olympic Games) will show a periodic increase or decrease. Relatively speaking, the media, the government and enterprises pay the lowest attention to leisure sports. The right of voice is naturally also the lowest. Take the American media, and their coverage of the four major American leagues, especially the Super Bowl and the NBA, is significantly higher than other sports.

The culture of sports discourse is mainly reflected as the "confrontation" and "entertainment" of sports. Conversarial projects, whether "control target type" or "control opponent type", will stimulate the enthusiasm of the media and audience because of the fierce competition among the athletes, thus attracting more investment from companies and the government. The discourse power ratio based on cultural variables is second only to economic factors. In contrast, the discourse group focusing on "control target" sports (such as basketball, tennis) seems to be larger than that focusing on "control opponent sex" sports (e. g. shooting, long jump). This is mainly because the former is mostly a group movement, which is much higher intense than the latter, which is mainly individual movement. Words about leisure sports are more common in "we media", such as jogging sports discussion, blog articles and so on.

The popularity of sports is determined by tradition, policy, and engagement. On the one hand, most Olympic events are accepted by ethnic groups around the world due to traditional and policy influences. On the other side, some Olympic events are quickly "forgotten" after the games due to different participation. On the contrary, it belongs to the "non-professional" sports, and the "non-competitive" leisure sports are more popular than the "competitive" sports. The latter is limited by hardware conditions and professionalism, and the penetration rate is not high, which also means that the discourse community is not large. Some sports with great regional characteristics, such as Chinese square dance, only play a certain part in the regional sports discourse. Therefore, from the perspective of nationality, the weight ratio of different

sports discourse is successively the global traditional projects-leisure sports events-emerging competition events-regional projects.

To sum up, the factors affecting the heavy voice of sports is economy first, culture later, and then to the degree of participation. Since discourse is present dynamically, and occurs in the context of "specific individuals, specific information, specific time and place" [17], one can never accurately determine the exact number of discourse, but can accept the fact that the amount of discourse varies in size. According to the discussion above, this article tries to give 10 to 1 weights to the 10 variables involved in the sociality of sports (see the left column of Table 2), and to calculate the possible discourse weight ratio of various sports. It should be stated that the assignment is not to give a precise proportion (which can only be calculated under a limited time period and context), but to reflect the differences in the amount of discourse, providing a feasible reference standard for the collection of various articles in the comprehensive corpus.

Table 2. Weight ratio of sports classification in sports discourse

Variables (weight value)	Project type						
	Category 1	Category 2	Category 3	Category 4	Category 5	Category 6	Category 7
Professionalism (10)		10		10			
Competition type (9)	9		9		9	9	
Same Court Contact (8)		8			8		
Net Dividing Contact (7)				7		7	
Stand-alone type (6)							6
With Non-equipment (5)			5				
Equipment (4)	4						
Globality (3)	3		3	3	3	3	3
Regional (2)		2					
Non-competition (1)							1
Total Value;	16	20	17	20	20	19	10
Weight ratio /%	13	17	14	16	16	16	8

The ratio of the obtained value and weight value of each sports category calculated by assigning values to variables of different sports categories can be regarded as the proportion of the relevant corpus of this category in the comprehensive Sports English corpus. For example, if the total corpus is 100 million words, the second category of professional sports (17%) should account for about 17 million words, while the seventh category of leisure sports (8%) accounts for about 8 million words. The larger the weight is, the higher the discourse output under this variable is, and the larger the proportion of corpus in the corpus is. In other words, the more popular a sport is, the more attention it attracts from the media, schools and the general public, and thus the more corpus it generates. Only by considering the imbalance of corpus distribution in the real world, can the actual situation of sports discourse be reflected in reference corpus more truly, so as to ensure the balance of corpus.

Similarly, discourse weight ratio needs to be discussed in different registers. There is also a natural imbalance in the proportion of utterance in different registers. Take people's reading as an example. No matter what their reading habits are, generally speaking, especially in the Internet age, people mostly read "recreational reading", such as reading news, comments, stories and advertisements. The second is "serious reading", mainly to acquire knowledge, such as textbooks, academic works, papers, etc. The third is "practical reading", which, as the name implies, can be used immediately after reading, such as reading the rules and arrangements of the event. From a marketing point of view, consumption determines output. The scale of corpus

output also tends to decrease from entertainment to seriousness to practicality. After considering the difference of discourse weight between sports classification and register classification, the balance of reference corpus can be guaranteed to a greater extent and the reality of sports discourse can be reflected more accurately.

3.2. Representative and Comparability of the Corpus

The principle of corpus representation is about the size of the corpus samples. In other words, how much of the corpus is the most appropriate? The ideal corpus is to bring all together into a super corpus. However, with language constantly being used by people, the idea can never be realized. To approach this goal, the surveillance corpus represents a constant language change, such as the NOW corpus, as the corpus increases regularly. But if the questions studied are specific, such as exploring the "adversarial" semantic feature [7] of the prepositional "against" [7]. Since the word has quite high frequency and critical values in the NBA event coverage corpus, only a smaller corpus (about 100,000 words) is needed to answer the concerned question. Therefore, the strength of the corpus representation is determined by the purpose of creating the corpus. For a comprehensive sports English corpus, on the basis of covering as many topics as possible and structural guaranteeing the principle of balance, the corpus should be as large as possible, large enough to include all sports, all discourse types, up to all specialized vocabulary.

The comparability of the corpus is specific for diachronic changes in the language. As stated in Section 3.1, a diachronic study of the sports English corpus is not the main objective of this study. Compared with general English corpora, specialized English corpus focuses more on their own synteny language characteristics, so the comparability of sports English corpus is mainly limited to the comparison between different domains, or the comparison of a specific corpus and comprehensive sports English corpus. If permitted, a regularly updated monitoring corpus can also be built to provide a corpus for the diachronic comparison study of sports discourse.

4. Conclusion

This paper provides a "three-dimensional model of social-language-quantitative" for the construction of a sports English reference corpus. Establish the classification of sports from the social dimension, the external factors of the corpus. From the language dimension, the intrinsic factors of the corpus, based on the story typology of systematic functional linguistics, the domain localization of various sports text is conducted, and the typical sports story type is identified. From the quantitative dimension, the equilibrium of sports English corpus is discussed, and the balance of reference corpus must reflect the discourse power weight ratio of language reality. The greater the discourse, the more corpus in the corpus and vice versa. The representativeness and comparability of the reference corpus can be solved by the construction of the monitoring corpus. In short, the social dimension is a table of the corpus, and the language dimension is the inside, while the quantification principle is the adhesive between them.

References

- [1] Chen Shan. English Professional Curriculum Reform and Setting in Institutions of Higher Learning [D]. Shanghai: Shanghai Foreign Studies University, 2012.
- [2] Chen Jiansheng, Tan Qingsong. Corpus, dedicated to the purpose of English and Sports English corpus construction [J]. Journal of Guangzhou Institute of Physical Education, 2009,29 (3): 111-115.
- [3] BAKER P. Glossary of Corpus Linguistics [M]. Edinburgh: Edinburgh University Press, 2006.
- [4] MCENERY T, HARDIE A. Corpus Linguistics :Method, Theory and Practice [M]. Cambridge :Cambridge University Press, 2011.

- [5] Lu Feng. On the Nature and Social Characterization of Sports [J]. Journal of Chengdu Sports College, 2001,27 (6); 10-13.
- [6] STEFANI R T. A taxonomy of sports rating systems [J]. IEEE Transactions on Systems,Man,and Cybernetics-Part A:Systems and Humans,1999,29(1): 116-120.
- [7] Chen Jiansheng. Common vocabulary is unusual-A corpus-driven study of the preposition "against" in sports event reports [C]. Chinese ESP Study, 2012,3 (2): 26-32.
- [8] SLACK T. The Commercialisation of Sport [M].Abingdon: Taylor & Francis Ltd,2004.
- [9] KACHRU BRAJ B. The Other Tongue: English Across Cultures [M]. Illinois : University of Illinois Press. 1992.
- [10] BIBER D. An analytical framework for register studies [C]//Sociolinguistic perspectives on register. New York,Oxford University Press.1994 :31-56.
- [11] DAVID YW LEE. Genres, registers, text types, domains,and styles :Clarifying the concepts and navigating a path through the BNC jungle [J].Language Learning & Technology,2001,5 (3) :37-72.
- [12] CHEN Jian-sheng. Construing Experience in tourism discourse : A corpus—based study of transitivity systems [D]. Hong Kong : The HongKong Polytechnic University, 2014.
- [13] HALLIDAY M A K,HASAN R. Language,Context,and Text: Aspects of Language in a Social Semiotic Perspective [M]. Oxford :Oxford University Press, 1989.
- [14] MATTHIESSEN C M I M, KASHYAP A K.The construal of space in different registers ; an exploratory study [J]. Language Science, 2014(45):1-27.
- [15] LEMKE J L. Ideology,intertextuality,and notion of register [C]// Systemic Perspectives on Discourse.New Jersey : Ablex, 1985:278.
- [16] SHAROFF S. In the garden and in the jungle :Comparing genres in the BNC and Internet[C]//Text,Speech and Language Technology. Dordrecht:Springer,2010:149-166.
- [17] CRESSWELL M J. Static semantics for dynamic discourse[J]. Linguistics and Philosophy, 2002, 25(5/6):545- 571.