Lecture Notes on Language and Literature (2024)
Clausius Scientific Press, Canada

DOI: 10.23977/langl.2024.070225
ISSN 2523-5869 Vol. 7 Num. 2

A comparative study of synonyms based on BNC corpus: A case study of IGNORE and NEGLECT

He Ying

Southwestern University of Finance and Economics, Chengdu, 610000, China

Keywords: BNC corpus; register; synonyms; collocation; semantic prosody

Abstract: In English, numerous words are considered synonymous, as dictionaries and thesauruses often provide identical definitions for them. This poses a significant challenge for English as a Foreign Language (EFL) learners, as they may perceive synonymous words as interchangeable, leading to potential ambiguity or awkwardness in their language use. However, the emergence of corpora and concordancing programs has introduced a new approach to investigating and learning synonymous words in English. This study specifically examines the collocational behavior and semantic prosody of two ostensibly synonymous verbs, "ignore" and "neglect," using the BNC Corpus. The findings reveal that these "equivalencies" can be misleading, as these words are typically used in distinct ways. The research suggests that learners may not have a comprehensive understanding of the collocation and semantic prosody of "ignore" and "neglect." Their tendencies to underuse or overuse certain grammatical forms and lexical patterns may be influenced by their native language or the registers of their English writing, or a combination of both factors. The paper concludes by discussing the implications of corpus-based studies for vocabulary teaching and learning more broadly.

1. Introduction

Second language instruction is designed to improve learners' vocabulary skills, which are key to improving skills such as writing, reading, and speaking in English. However, the study of words has only just begun in modern times, and the word "Lexis" has only a relatively short history in English, and it was not until about the 1950s that it began to be used in British English. Lexical competence has been identified as one of the most significant predicators to general language ability. It is an aspect of both linguistic competence and communicative competence and refers to the ability to produce and understand the words of a language. But it is also admitted by most learners to be a big challenge in language learning. With the advent of corpora and concordancing programs, a new view of teaching and learning vocabulary has emerged. Compared to traditional tools for learning vocabularies like dictionaries, using a corpus to learn vocabulary has certain advantages. On the one hand, a corpus offers real language examples, making the language learning environment more authentic. On the other hand, unlike dictionaries that mainly provide sentences for word meanings, a corpus shows examples (concordances) of target words in a context, either in a sentence or in a larger text. It can even display the whole text where the word is used. In addition, these examples are usually shown in a way called KWIC (keyword-in-context), which helps learners notice how words are used and the

grammar around them. In this sense, learners may learn not only word meanings but also word usage in specific contexts. Synonyms are a major difficulty in L2 vocabulary teaching and because they have great similarities in structure, word form, word meaning and semantics.

2. Literature review

2.1. Research status at home and abroad

The study of words has always been one of the core of Firth's linguistics, and it can be said that lexicology is mainly rooted in Firth's theory of meaning. Later, his students, Sinclair and Halliday, among others, inherited and developed his ideas, proposing important concepts such as semantic prosody. The study of lexicology is inseparable from corpora, the earliest corpus established is the Brown corpus, and later COCA, COBUILD, etc., and the BNC used in this paper is also one of them. Foreign scholars, such as Biber have used corpora to explore the differences in the context and register of synonym[1] In his research "Register as a predictor of linguistic variation", he employs a methodological approach rooted in corpus linguistics to examine the relationship between register variation and linguistic features. Biber selects extensive corpora, such as the Corpus of Contemporary American English (COCA), to conduct his analyses. Employing quantitative linguistic methods, he scrutinizes large datasets to discern the frequency and variations in the usage of specific words, phrases, or grammatical structures across different registers. For instance, Biber investigates the usage of vocabulary in diverse contexts, spanning academic articles, news reports, and spoken conversations, aiming to identify variations based on discourse types and social contexts. Through this methodology, he unveils the nuanced adaptability of language, providing profound insights into how linguistic elements dynamically respond to contextual influences. Implications for linguistic lexical research, which can reveal the flexibility and adaptability of language in different contexts, and help us better understand the diversity of languages. In addition, through the analysis of large-scale corpora, we can understand the patterns of language use more comprehensively, and provide useful information for language teaching and language policy making.

There are several main aspects of domestic research: First, the study of synonyms based on the corpus of the native language, including conjunction, semantic rhyme, etc. For example, Xie Zhuojun(2020) [2]used corpora to distinguish the difference between quite and rather and Wang Jia(2022)[3] made a comparative study of the semantic prosophies of the words effect and result. The second is to compare the corpus of native languages with the corpora of Chinese scholars to infer the reasons for the inappropriate use of words by Chinese students. So we can study lexicology based on corpora.

Synonyms refer to a group of words that contain the same or nearly identical essential conceptual meaning. The traditional dictionary-based discrimination method lacks scientific and limitations, while the rise of corpus-based research has brought a more scientific, objective and comprehensive research method to synonym discrimination. Synonyms, typical collocations, and grammatical patterns contained within them cannot be changed at will. Based on previous researches, this paper will identify synonyms from the following five dimensions, primarily word-based approaches

2.2. Frequency

One of the most important concepts in corpus linguistics is frequency, which refers to the frequency of a word or phrase in a corpus per million words, also known as normalized frequency. There are often subtle differences between synonyms, which means that they often exhibit different distribution characteristics in different registers. Therefore, the frequency difference of synonyms in the corpus or in different registers can reflect the differences in their meanings and help English

learners learn synonyms better. The frequency distribution can reflect the probability properties of the language system to a certain extent. (Halliday, 2014)[4]

2.3. Register

Registers refer to a language variant that people have in actual language activities out of the need for communication, or because of their different occupations and interests, as well as because of the different situations in which their words occur, the objects they speak, the places and the topics, which are reflected in the different styles and styles of language in the language. To put it simply, domain is the environment of language use, and any language use is affected by the register, and the use of language in different registers presents different characteristics. Register plays an important role in the discrimination of synonyms. The importance of register for vocabulary is well-established from corpus research (Kennedy,2014)[5]. To take an easy example, the pronouns I and you are among the most frequent words in the spoken LondonLund Corpus, but considerably less frequent in the written Brown Corpus. ELT dictionaries like the Longman Dictionary of Contemporary English provide detailed information of this type, explicitly identifying the most frequent words in speech versus the most frequent words in writing. Many studies have investigated the preferred collocates of specific target words. For example, Sinclair (1991)[6] describes the uses of phrasal verbs with set and collocates of the word back. Hunston (2002)[7] discusses the phraseological patterns of several target words, such as recipe, initiative, condemn, suggestion, point, gaze, leak, and shoulder. Partington (2004) [8], in a book-length treatment of collocation, provides detailed descriptions of the phraseological patterns for sheer, pure, complete, absolute, correct, absolutely, completely, entirely. These studies provide detailed descriptions of the collocations and preferred uses of a specific target word, and further illustrate how supposed synonyms are not in fact identical in meaning or use when considered from this perspective. However, these studies are typical in that they include no mention of register or the possibility of different word uses in different registers. By searching the British Contemporary Corpus (BNC), the ignore and neglect frequency and frequency results were obtained. Since there are differences in the intrinsic meaning of synonyms, that is, they will show different distribution characteristics in different registers, the frequency difference in different registers can help students distinguish synonyms and master authentic vocabulary expressions.

2.4. MI (Mutual Information)

"Hunston (2002) [7] defines collocation as the 'tendency of words to be biased in the way they co-occur.' In the pursuit of understanding the crucial collocates associated with a word, two key metrics come into play: mutual information and T-score. This paper delves into the significance of Mutual Information (MI) scores.

Mutual Information, henceforth referred to as MI score, is used to calculate the number or actual occurrences of a word against the number of times that word was predicted to occur. Hunston(2002) [7] says that '…MI score measures the amount of non-randomness present when two words occur.' Hunston and Laviosa (2000)[9], state that this gives a more accurate idea of the relationship between two words. They go on to say that MI score assesses the importance of a collocation and that it shows a clearer picture of the relationship between words than that given by a simple collocation list alone. It is a measurement of two-way attraction. Walter (2010:435)[10] contributes to this discourse by stating that the infrequent co-occurrence of a word with another implies that this collocation is unlikely to happen by chance. Nevertheless, Baker (2023) [11] introduces a caveat, noting that MI scores may overly emphasize words occurring rarely in a text, potentially yielding somewhat misleading results. The accuracy and practicality of these results remain ambiguous. Hunston (2002) [7] recommends considering MI scores of 3 or higher as significant. This paper aims to not only discuss

MI scores but also to validate the importance of specific collocations while calculating MI scores."

2.5. Collocation

In contemporary vocabulary learning, two concepts gain prominence: the scope of vocabulary knowledge and the depth of vocabulary knowledge (Pu 2003:439)[12].The former notion refers to the number of words one has known, while the latter one deals with the extent to which one has grasped the usage of words (Pu, 2003:439). Apparently, the analysis of synonymous words is to help students with the depth of vocabulary knowledge. By the depth of vocabulary knowledge, there are two components: (1) mastery of the core grammatical patterns of words; (2) mastery of the typical combinations of words (Pu 2003:439)[12]. These two concepts are explained in terms of "colligation" and "collocation". This paper will specifically explore the collocation of these two words. When learners acquire vocabulary in isolation, deviations may occur. Thus, it becomes imperative to establish meaningful rules based on the most common collocations of a word. Understanding a vocabulary necessitates awareness of its typical pairings. Given that collocation knowledge is crucial for the correct and fluent use of language, it should assume a central role in lexical research.

One component of the depth of vocabulary knowledge, collocation, has been studied for at least fifty years. Hunston and Laviosa (2000)[9], state that collocation is the propensity for words to occur near each other in a text. In other words, they co-occur, or they are co-located. However, they also point out that just because two words frequently occur near each other, this does not necessarily mean that there is a high significance to this co-occurrence. Firth, a pioneer in the study of collocation, defined it as 'actual words in habitual company' (Firth, 1957: 99) [13]. By searching with node words as the center, the number of words displayed on the left and right forms the miniature context of the node words.

2.6. Semantic prosody

Transitioning from linguistic forms to meanings, recent corpus-based studies reveal that certain grammatical forms inherently harbor semantic relations between words and their collocates, as well as among the collocates themselves (Stubbs 2002: 225).[14] This intricate interplay gives rise to what is termed 'semantic prosody' (Louw 1993: 157).[15] Louw (1993:157)[15] defines semantic prosody as 'a form of meaning established through the proximity of a consistent series of collocates, whose primary function is to express the speaker's or writer's attitude or evaluation'. In essence, semantic prosody manifests in three categories: positive, neutral, and negative (Stubbs,1996). [16] A positive semantic prosody indicates that node words exist in a positive semantic atmosphere, while a negative semantic prosody implies the opposite, characterized by overtly negative nuances. Neutral semantic prosody falls somewhere in between, lacking the extreme tones of positivity or negativity. This exploration of semantic prosody serves as a critical lens through which to understand the nuanced and evaluative dimensions of language use, shedding light on how words acquire connotations based on their consistent associations in context."

2.7. The Gap of research method about synonyms discrimination used in the current mode of vocabulary teaching.

Based on the multiple research dimensions of previous scholars, this paper analyzes the differences between a group of synonymous verbs IGNORE and NEGLECT from five dimensions. The pair of synonym verbs selected in this paper is just one example of a pair of synonyms, and educators can take the same dimensions to study the correct use of more synonyms. If educators want to conduct a comprehensive study of vocabulary, including grammar, pronunciation, context, etc., they can use a

combination of foreign corpus tools and domestic corpus tools, and a large number of corpus is required as the basis for research, and the process is complex and time-consuming. So, educators can study the five dimensions of this paper in the current routine vocabulary teaching, and this method can be applied to the current mode of vocabulary teaching.

3. Research problems

This study will address the following research problems:

Frequency Variation: the frequency value of IGNORE and NEGLCET in BNC.

Register Influence: What role does register play in the discrimination of synonyms, and how can corpus-based research contribute to a better understanding of synonym usage in varied contexts?

MI Score: The degree to which the two synonyms IGNORE and NEGLECT are paired with their collocations.

Collocation Knowledge Impact: To what extent does collocation knowledge contribute to the depth of vocabulary knowledge, and how can this be applied in language education?

Semantic Prosody Influence: Through the semantic prosody analysis of collocations, the attributes of IGNORE and NEGLECT are judged, and how to use them in the correct context.

4. Methodology

The research will employ a corpus-based analysis, utilizing existing corpora such as the British National Corpus (BNC). The selected corpora will be queried to extract relevant data for each dimension, and statistical analyses will be applied to draw meaningful conclusions.

The native language corpus that will be used in this study is the written language corpus of the British National Corpus (BNC), which is a large corpus jointly developed by Oxford Press, Longman Publishing Company and Lancaster University English Computer Center, with a capacity of 100 million words, and the written language part accounts for 90% of the total library capacity, and the collection of a wide range of corpus, including national newspapers, magazines, novels and university papers. In addition, the corpus also has its own corpus retrieval system, which can obtain the index rows where the words are located and their significant collocations online.

First, enter IGNORE and NEGLECT on the search page, and then query their frequency and record datas. Secondly, to understand the register distribution, the author selects the DISTRIBUTION interface, and a variety of CATEGORIES will appear, at which time, record the frequency and frequency per of the distribution of these two words in different registers. Following Krishnamurthy's recommended approach, the corpus search will focus on collocations within a range of 5 words on the left and right sides of the node words for observation. Collocation strength will be determined by Mutual Information (MI) values between collocation words and node words. The characteristics of collocation words will be analyzed, with a particular emphasis on observing MI values. A threshold MI value of 3 or higher will be considered indicative of significant collocation intensity, establishing a positive correlation between MI value and collocation intensity. In this study, "ignore" and "neglect" will serve as the node words. Collocations within 5 words on the left and right of these nodes will be retrieved, capturing the co-occurrence frequency and MI value of collocation intensity for each synonym pair. Non-word symbols will be excluded based on MI values, and the top 10 collocations with the highest frequency will be selected as the primary focus of investigation. This refined methodology aims to provide a comprehensive understanding of synonym discrimination by systematically analyzing collocations within a well-defined linguistic context. The focus on MI values ensures the selection of collocations with substantial significance, contributing to the depth and accuracy of the study's findings.

5. Differences between ignore and neglect in the BNC corpus

5.1. Differences in the frequency distribution of ignore and neglect in different registers of the corpus

By searching the British Contemporary Corpus (BNC), the ignore and neglect frequency and frequency results were obtained. Frequency is one of the most important concepts in corpus linguistics, and the distribution of a word or phrase can be seen by calculating the frequency of a word or phrase in a corpus per million words (also known as normalized frequency). Since there are differences in the intrinsic meaning of synonyms, that is, they will show different distribution characteristics in different registers, the frequency difference in different registers can help students distinguish synonyms and master authentic vocabulary expressions. (Table 1)

Table 1: Statistics on the frequency of ignore and neglect occurrences in the BNC corpus

Register	ignore		neglect	
	frequency	Frequency per million words	frequency	Frequency per million words
Spoken	2195	24.97	1160	13.2
Written	208	19.98	26	2.5
Academic prose	489	30.99	379	24.02
Fiction and verse	484	29.98	100	6.19
Newspaper	179	19.02	82	8.71
Non-academic prose and biography	553	22.87	363	15.01

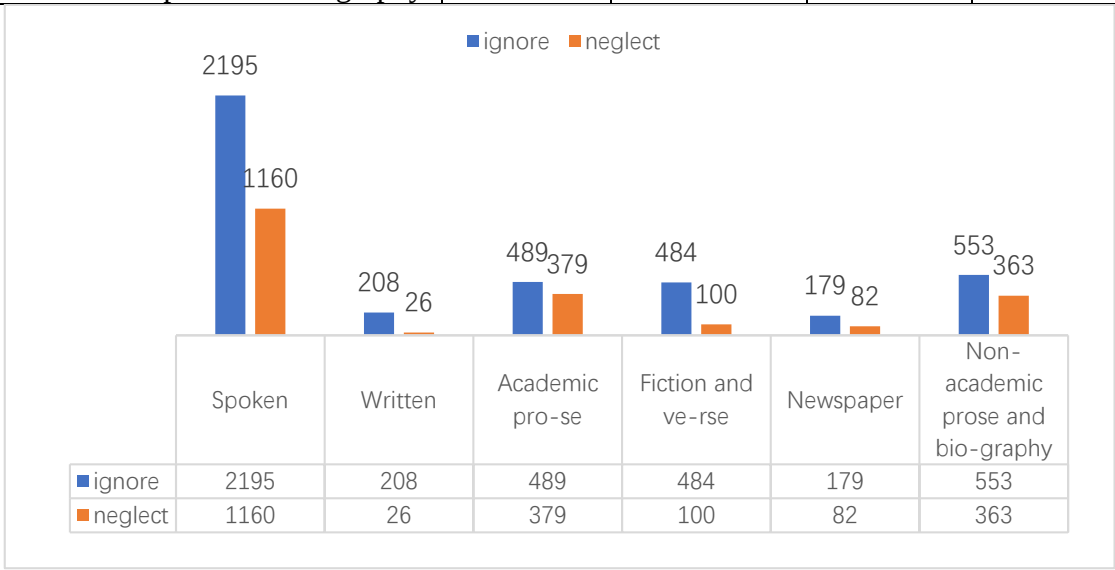

Figure 1: Frequency distribution of ignore and neglect in the BNC corpus

In addition, the frequency histogram of ignore and neglect in each register will clearly show the difference in the distribution of the two in different registers. The above data shows that the frequencies of ignore and neglect in the BNC corpus are quite different. In general, ignore is used several times more frequently than neglect. In specific registers, ignore is used most frequently in the Spoken register, and is used equally frequently in other registers. Neglect is also the most frequently used in the Spoken register, and less frequently in Written and newspapers. Regardless of the register, ignore is used more frequently than neglect. It's worth noting that ignore appears 8 times more often

than neglect in the Written register. By comparing the frequency distribution of these two words in the BNC corpus and their frequency in different registers, it is concluded that ignore is used more frequently than neglect. In the actual use process, learners can use the corpus to analyze the frequency of using of synonyms in various registers. Select the corresponding vocabulary according to the register to avoid the misuse of words caused by mechanical memory. (Fig 1)

5.2. Collocation of ignore and neglect

In this paper, the author will discuss about the collocation of these two words. The collocation strength is calculated by the MI values of the collocation words and node words, and the characteristics of the collocation words are analyzed. We can observe the MI value, which is the Mutual Information Value. When the MI value is greater than 3, it is considered that the collocation has a significant collocation intensity, and the MI value and collocation intensity are positively correlated. In this paper, ignore and neglect were used as node words, and the collocations within 5 words on the left and right of the two were retrieved, and the co-occurrence frequency and MI value of collocation intensity of each collocation and node word of this group of synonyms were obtained. According to the MI value, non-word symbols were excluded, and the top 10 collocations with the highest frequency were selected as the research object, and it was found that the collocations of ignores were mainly verbs while neglect were some derogatory adjectives. (Table 2)

However, neglect cannot be classified as a negative semantic rhyme type, and it is necessary to examine the extended context in which the collocation is located, because it is inaccurate to discuss the semantic prosody of words without context.

Table 2: Glossary of significant collocations of ignore/neglect in BNC

BNC				
	ignore		neglect	
Rank	collocate	MI value	collocate	MI value
1	peril	7.9398	wilful	9.2487
2	chose	6.3898	carelessness	8.3483
3	afford	6.0184	mismanagement	8.1021
4	chooses	5.9589	benign	7.6735
5	warnings	5.7764	abuse	7.1849
6	ignore	5.7644	cruelty	6.9718
7	pretended	5.7228	default	6.8055
8	fascism	5.4217	decay	6.7816
9	foolish	5.2622	indifference	6.5819
10	tended	5.258	attributable	6.3435

5.3. The semantic prosody of ignore and neglect

From the perspective of the semantics of collocations, the terms that are significantly matched with neglect can be divided into the following categories: (1) Adjectives such as wilful are negative adjectives, benign have a positive meaning, and attributable belong to neutral words; (2) Nouns with negative meanings: carelessness, mismanagement, cruelty, indifference, abuse;(3) The verbs decay and default are both negative meanings. So, neglect has negative semantic prosody. The semantic categories of the collocations of ignore are roughly the same as those of neglect, and are mainly divided into the following categories:(1) Verbs include chose, afford, chooses, ignore, pretended, tended and most of which tend to be neutral;(2) Nouns with negative meanings, such as warnings,

fascism, peril;(3) derogatory adjectives, such as foolish. From the above analysis, it can be seen that both ignore and neglect are more inclined to be paired with negative words, and it can be seen from Table 2 that the intensity of neglect and negative words is higher than that of ignore (the MI values of wilful and carelessness are 9.2487 and 8.3483, respectively, which are higher than the MI values of ignoble and negative words, and the higher the MI value, the higher the matching intensity), indicating that learners are differentiating between ignore and neglect tends to be applied to negative contexts. Thus, ignore is a mixed semantic prosody with neutral and negative meanings. It is precisely because the collocation of ignore has both a general neutral collocation and a negative connotation of the word collocation that the semantic prosody of the word shows an intricate situation. (Table 3)

Table 3: Ignore/neglect table of semantic prosody types in BNC

verbs	positive	negative	neutral	Semantic prosody
ignore	10%	30%	60%	neutral
neglect	10%	80%	10%	negative

6. Implication and Conclusion

This research aims to provide valuable insights into synonym discrimination, exploring the multifaceted dimensions that influence the comprehension and utilization of synonyms. Through a corpus-based analysis, the study delves into the collocational behavior and semantic prosody of seemingly synonymous verbs, IGNORE and NEGLECT, elucidating their nuanced distinctions. The findings emphasize the complexity of synonym usage, revealing that seemingly synonymous words are not interchangeable in certain contexts due to distinct collocational patterns and semantic prosody. This nuanced understanding underscores the significance of considering both grammatical properties and lexical features in vocabulary instruction.

The study advocates for a comprehensive approach to vocabulary teaching and learning, encompassing pronunciation, meaning, frequency, field of application, collocations, and semantic prosody. Language corpora emerge as crucial resources, providing authentic language materials and aiding learners in discerning proper and idiomatic word usage.

On the other hand, it's essential to acknowledge potential limitations in this research. The representativeness of the chosen corpora and inherent biases in corpus data may impact the generalizability of findings. Additionally, the focus on English synonyms limits direct applicability to other languages.

This research holds significance for language educators, learners, and researchers by offering nuanced insights into synonym discrimination. The findings may inform language teaching strategies, providing practical implications for distinguishing synonyms effectively. Moreover, the study contributes to corpus linguistics methodologies, advancing our understanding of how learners navigate the complexities of synonym usage in different linguistic contexts.

References

[1] Biber, D. Register as a predictor of linguistic variation[J]. Corpus linguistics and linguistic theory, 2012, 8(1): 9-37.
[2] Xie Zhuojun. Research on synonym analysis based on BNC corpus - taking quite and rather as examples[J]. English Square: Academic Research, 2020, (1): 2.
[3] Wang Jia & Zhang Hong. Corpus-based study on the acquisition of synonym semantic prosody by Chinese non-English major college students—taking effect and result as examples[J]. Journal of Jingchu Institute of Technology, 2022, (05): 89-96.
[4] Halliday, M. A. Corpus studies and probabilistic grammar[J]. In English corpus linguistics Routledge, 2014: 42-55.
[5] Kennedy, G. An introduction to corpus linguistics[J]. Routledge, 2014.
[6] Sinclair, J. Corpus, concordance[J]. collocation, 1991.

[7] Hunston, S. 'Corpora in Applied Linguistics.'Cambridge University Press[J]. Cambridge, 2002.
[8] Partington, A. "Utterly content in each other's company": Semantic prosody and semantic preference[J]. International journal of corpus linguistics, 2004, 9(1): 131-156.
[9] Hunston, S., & Laviosa, S. (2000). Corpus linguistics. In Corpus Linguistics (pp. 1-177).
[10] Walter, E. Using corpora to write dictionaries[J]. The Routledge handbook of corpus linguistics, 2010: 428-443.
[11] Baker, P. (2023). Using corpora in discourse analysis. Bloomsbury Publishing.
[12] Pu, J. Colligation, collocation, and chunk in ESL vocabulary teaching and learning[J]. Foreign Language Teaching and Research, 2003, 35(6): 438-445.
[13] Widdowson, H. Jr firth, papers in linguistics [J]. International Journal of Applied Linguistics, 1957, 17(3): 402-413.
[14] Stubbs, M. Two quantitative methods of studying phraseology in English[J]. International Journal of Corpus Linguistics, 2002, 7(2): 215-244.
[15] Louw, B. Irony in the text or insincerity in the writer? The diagnostic potential of semantic prosodies[J]. Text and technology: In honour of John Sinclair, 1993: 157, 176.
[16] Stubbs, M. Text and corpus analysis: Computer-assisted studies of language and culture[J]. Oxford: Blackwell, 1996: 158.