This is an Accepted Manuscript of an article published by Taylor & Francis Group in *Journal of quantitative linguistics,* published online: 09 Aug 2010, available online at: https://www.tandfonline.com/doi/abs/10.1076/jqul.9.1.35.8483

# Quantifying the Semantic Contribution of Particles Submitted to Linguistics

Ramon Ferrer Cancho*
Complex Systems Research Group, FEN, UPC
Campus Nord, B4-B5, Barcelona 08034 SPAIN

Francesc Reina
Secció de lingüística
Department de Filologia
Facultat de Filologia
Universitat de Barcelona
Barcelona, SPAIN

December 29, 2000

**Abstract**

Certain natural languages word types: conjunctions, articles, prepositions and some verbs have a very low or very grammatically marked semantic contribution. They are usually named functional categories or relational items. Recently, the possibility of considering prepositions as simple parametrical variations of semantic features instead of categorial features or as the irrelevance of such categorial features has been pointed out. The discussion about such particles has been and is still widespread and controversial. Nonetheless, there is no quantitative evidence of such semantic weakness and no satisfactory evidence against the coexistence of categorial requirements and the fragility of the semantic aspects. This study aims to quantify the semantic contribution of particles and presents some corpora-based results for English that suggest that such weakness and its relational uncertainty come from the categorial irrelevance mentioned before.

## 1 Introduction and goals

The explanatory role of categorial features of particles like prepositions undergoes remarkable theoretical dificulties. Not only the classical description of this

*To whom correspondence should be addressed. e-mail: ramon@complex.upc.es Phone: +34 93 4017056. FAX: +34 93 4017100.

class of words shows clear contradictions when speaking about functional and semantic roles, but also certain empirical phenomena -relevant in different natural languages- are in evident descriptive oposition. A lot of linguists discuss and propose certain explanatory reductions of the prepositional category to either linguistic features [19], reanalysis with another category (nouns, verbs, adverbs or conjuntions) [24] or incorporation processes [2]. Nonetheless, a lot of prevailing considerations insist on the homogeneous kind of their categorial behaviour as if they were nouns, verbs or adjectives. The starting point of such views comes from their phonological and morphological independence. Such grounds are dubious since they rely on an interlinguisticaly parametrical epiphenomenon.

The case markedness (accusative, dative,...) is apparently the evidence of the categorial consistence of prepositions regarding hierarchy construction and syntactic relation. The named objects of the predicates are introduced in most languages from marks, either morfologically or with a strict position. Phrase structure grammar of the generativist tradition started its inquiries for English with Jackendoff [15] and his proposal of structural endocentricity for prepositions. He was followed by Van Riemsdijk [23], Emonds [7], Chomsky [4] and Kayne [17]. These linguists started to show certain special features that are the basis of the most recent proposals.

Prepositional semantics has been studied in turn from a double view. Cognitive linguistics assumes certain rudiments of human cognition from the Gestalt's psychological tradition (e.g. the distinction between figure and ground) in the description of certain semantic aspects of the expression of space that affects prepositions. Prepositional meaning is related to the semantic competence of human beings in processing and interpreting concepts such as space.

The second view acknowledges that prepositions mark the arguments of the linguistic predicates. From the first works by Gruber [12] and Fillmore [8] to the conceptual structures by Jackendoff [16], a lot of generative contributions about the meaning have considered that this function or role is visible to the syntax. This is the case of the lexical developments starting on the X"-bar convention that arose during the eighties, such as Baker's UTAH hypothesis [2] and the lexical syntax by Hale & Keyser [18].

## 1.1 Theoretical foundations

In our work the following questions coming from dependency grammars and information theory are assumed. The methodological assumptions of the dependency grammars can be found in Melčuk [22] and Fraser [9] and certain versions of the Word Grammar by Hudson [14]. Specifically, grammatical relations are *primitives* of dependencies between words. There are no constituents. Relations take place directly and only between words. Words attract, link and relate themselves. The structure of a sentence is a graph (a tree) whose links are pairs of words. We do not take into account the direction of the link. These minimal assumptions (straight links between basic units and undirected links) have been successfully used for discovering relevant properties (scaling and/or small-world) in other systems having a well defined graph structure such as the

WWW, neural networks, collaboration nets and power grid networks [25, 1].

Information theory provides different ways of quantifying the semantic contribution of events (words). The first one and the simplest is self-information. Roughly speaking, the most (self-)informative words are the less frequent ones. The quantification of the information contained in links between words is acomplished by mutual information (MI). The value of MI between a pair of words keeps a more subtle relation between the frequence of the intervening words. Formulae and some intuitions about the behaviour of MI are given in subsection 2.1.

### 1.2 General goal and hypothesis

The general goal of our work is to quantify the semantic contribution of prepositions mainly from a representative number of triples of cooccurence $(x, y, z)$, where $y$ is a preposition (the election of y as the preposition is merely notational). For instance, the sentence

*"The UAW is seeking a hearing by the full 14-judge panel".*

provides the triple $(x, y, z) = (hearing, by, panel)$. 2228 prepositional triples where extracted from the Wall Street Journal Corpus. Our hypothesis is that the heaviest link of a triple (weight measured in terms of average mutual information) is $(x, z)$ because $y$ is not very significant.

## 2 Quantification

### 2.1 Approaches

The self-information (semantic contribution) of a word whose probability (frequence) is $p$ is $-logp$. From the self-information point of view, particles are the least informative words because they are the most frequent words. Moreover, Zipf's studies [28] revealed that their frequence is significantly higher than the rest of the words (several orders of magnitude in a lot of cases). The self-information of a word does not allow to infer how meaningful (informative) the links between such word and others are.

Mutual information provides a way of measuring the strength of the correlation between two words (in other words, how reliable its link is) and has been succesfully used for measuring correlations between elements in genomic sequences [20, 6]) and texts [6, 21] and in syntactic disambiguation [10].

The *average mutual information* (AMI) between two words $w_i$ and $w_j$ (hereafter simply 'i' and 'j') is $f(p_i, p_j, p_{ij}) = p_{ij} I_{ij}$, where $I_{ij} = log\frac{p_{ij}}{p_i p_j}$ is the mutual information, $p_i$ is the probability that $i$ participates in a link ($p_j$ is the same for $j$) and $p_{ij}$ is the probability that both words participate in the same link (joint probability). Notice that $I_{ij} = I_{ji}$ (symmetry), which is consistent with the assumed undirectedness of links.

Once $p_i$,$p_j$ and $p_{ij}$ are known, the calculation of $f$ is straightforward. Several methods have been studied for estimating such values [26, 27, 10]. Subsection 2.2 and Section 3 show, respectively, the method we use for determining the links in prepositional triples of cooccurrence and the results obtained.

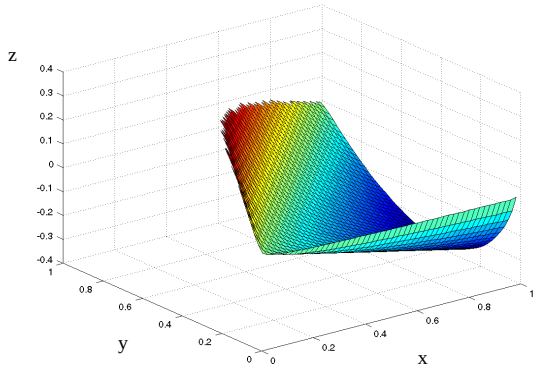


Figure 1: Average mutual information ($z = p_{ij} log \frac{p_{ij}}{p_i p_j}$) as a function of the product of the probabilities of the pair of events ($x = p_i p_j$) and the joint probability ($y = p_{ij}$).

Some general predictions can be done before estimating the probabilities for a particular kind of particle. Figure 1 plots de value of $F(x, y)$, the AMI as a function of $x = p_i p_j$ and $y = p_{ij}$. Values of $p_{ij} > min(p_i, p_j)$ are omitted because they do not make sense. It can be seen that $F$ vanishes while $p_i p_j$ is increased and $p_{ij}$ is constant. On the other hand, an increase in $p_{ij}$, while $p_{ij}$ is constant, turns into an increase or a decrease of $F$ depending on the side of the valley taken as a starting point. The first consequence is that the strength $f_1$ of a link with a particle will be smaller than that of a link without a particle, $f_2$, if the joint probabilities of both links are the same. A formal proof is given in Appendix A. If the joint probabilities of both links are different, we cannot guarantee that $f_1 < f_2$. Notwithstanding, the valley sinks as $x$ grows,

so the range of values that provide an AMI greater than a certain cutting value gets narrower. The results obtained with real data fit this observation. Some analytical properties of the valley are shown in Appendix A.

### 2.2 Estimation of the probabilities and calculation of AMI

The estimation of the probabilities has been performed in different contexts. [26, 27] used them for calculating the strength of the association and thus determining the links between words in sentences. Furugori et al. [10] used them for determining the strength of the associacions between three cooccurring syntactic objects.

Let $\hat{p}_{ij}$ and $\hat{p}_{ij}$ be the estimated values of $p_{ij}$ and $p_i$. $\hat{p}_{ij}$ is defined as the probability of coocurrence of the $i$-th and $j$-th word in the same triple and $p_i$ is the probability of occurrence of the $i$-th word, according to [10] and one of the strategies used in [26].

Furugori et al. [10] pointed out the difficulties encountered in the calculation of the mutual information due to data sparseness and used word class occurrences to overcome such limitation. On the other hand, Yuret used only words with great success, despite of the obvious limitations of his approach. It might be thought that we also should use word classes. There exist classifications for content words but, what about prepositions? Providing a classification of prepositions would imply the acknowledgement of a certain status of prepositions before being able to state which one is the most appropiate. After having presented our results for our set of triples we will be ready for providing quantitatively-supported criticisms against classes for prepositions. The remaining question is whether to use at least classes for the partners of the prepositions. We are not aimed at checking how many links are well-detected because prepositional links are controversial, there are no test corpus for checking the results and again, it would imply to take certain assumptions beforehand without quantitative support. Classes for partners would improve the quality of the links detected for the most sparse cooccurences (Section 3 lists some links that would obviously improved with these approach) but our main argument - the direct relationship between low self-information and low average mutual information - would be still valid.

## 3 Results

We found that the pair x-z had almost the maximum weight (AMI) of the triple (x-z was nearly always *dominant*). The proportion of links in which x-z, x-y and y-z were dominant was 0.92, 0.023 and 0.025, respectively. A remaining 0.03 comprises the triples without unique dominant edge. The pair of links $\{(x,z),(x,y)\}$, $\{(x,z),(y,z)\}$ and $\{(x,y),(y,z)\}$ are named trees I, II and III, respectively. One way of determining the tree structure of the triple consists of choosing the pair of links whose sum of weights is maximum (as in [26, 27]). The proportion of trees I/II is the same as that of triples where $(x,z)$ is

dominant (trivial proof *ad absurdum*). Therefore, prepositions can be considered to have always linking degree equal to one in triples and thus in sentences. If prepositions have a degree equal to one in the trees of sentences, they are vertices that can be removed without disconnecting the tree. Words that disconnect the tree when removed are the key words of the sentence (such vertices of a graph are known as articulation vertices or cut-vertices in graph theory [3, 13]). The robustness of the results obtained is discussed in Appendix B.

During the extraction of the set of triples, we found problematic structures which we did not include: conjunctions or particles with a certain degree of grammaticalization and compound prepositions. The number of such structures was 49. Their presence did not allow us to determine the components of the triple or the distributional behaviour:

- Dubious conjunctions or special particle (30 years ago, We're talking about years ago before, a long way out, for example, for instance).
- Compound prepositions (according to Brooke, they blip down because of recent rises,...has annual revenue of about 370 million,its internal reorganization plan at about 2 billion, on the other hand).

# 4 Conclusions

We have shown that prepositions

- Tend to have weaker links. As a matter of fact, the link $(x, z)$ is the heaviest in prepositional triples. Self-information and average mutual information are coupled.
- Are (mostly) single linked (they have degree equal to 1) in tree representations of the relations between words.

# 5 Prospects

The most promising prospect will spring from typological contrastive estudy between languages. As Greenberg's universal linguistics proposal points out [11], the number of prepositional relations is very similar among most of human languages: locative and temporal deixis, oblicual cases like dative, benefactive and applicative. Notwithstanding, why does the number and the function of the lexical class that accomplish them vary? If the promise is valid and powerful, it must be of empirical kind. Linguistic theory will be able to ground its inquiries not on categorial supositions, having more or less degree of irrelevance or undeterminism, but on other class of supositions. As we mentioned, work on smaller items (features) might lead to more satisfactory conclusions. To this effect and closer to our work, the relations between x-y and y-z, that is to say, the constitution or the dependence between the preposition and its head(s) will be able to offer new views about the old syntax. The later developtment

of Transformational Grammar known as minimalist program [5], is working on certain simplifications and reductions of the syntactic objects. The simplest operations (for instance, merge and move) that constitute such objects are, may be, a theoretical prospect that can solve facts of the kind treated here, since it does not assume prior rules, conditions and restrictions.

# A Slopes and critical points of the average mutual information

Let $p_{ij}$ be the normalized joint probability of the events $i$ and $j$ and $p_i$ and $p_j$ the normalized probabilities of such events. Let $F(x,y) = f(p_i, p_j, p_{ij}) = p_{ij} log \frac{p_{ij}}{p_{ij}}$ where $x = p_i p_j$ and $y = p_{ij}$. The slope of $F$ while varying $x$ while $y$ is kept constant is given by

$$\frac{dF(x,y)}{dx} = -y/x$$

which is always lower or equal than zero because $x, y \in [0,1]$. An increase in $x$, i.e. an increase in $p_i$ and/or $p_j$, while $y$ is constant always leads to a decrease of $F$. The only critical point, $(x^*, y^*) = (x, 0)$ is not interesting for our purposes because $F(x^*, y^*) = 0$. Notice that $F(x, 0) = 0$ regardless of the value of $x$ because $x$ aproaches faster to 0 than $log \frac{y}{x}$.

The rect $y = \frac{x}{a}$ contains the projection on the x-y plane of the deepest points of the valley of $F$. Equation

$$\frac{dF(x,y)}{dy} = log \frac{y}{x} + 1 = 0$$

provides the critical points of $F(x,y)$ while $x$ is constant and $y$ variable. The critical points can be rewritten as $y = \frac{x}{a}$, where $a$ is the base of the logarithm. The second derivative $\frac{d^2(F(x,y))}{d^2 y} = \frac{x}{y}$ is always positive since $x, y \in [0,1]$. Thus, the rect is the projection of a valley.

# B Robustness of the learning procedure

We need to guarantee that the weights computed capture significant information. The transinformation of the entire set of triples, $I(S)$, will be defined as

$$\sum_{i,j} p_{ij} log \frac{p_{ij}}{p_i p_j}$$

where $i$ and $j$ are pairs of words in $S$. Being $I(S) \geq 0$, we have to guarantee that the information captured is significantly far from 0.

Two other control sets $S_r$ and $S_f$ were built. Let $S = (S_1, S_2, S_3)$ be a $n \times 3$ matrix whose rows are the triples being considered and $S_i$ the $i-th$ column of the matrix. Let $\Pi(x)$ be the permutation of a vector $x$. Let $concat(x,y)$ be a vector which is the concatenation of the vectors $x$ and $y$. Let $x = (x_1, ..., x_i, ..., x_n)$, $i \leq j$ and $ext(x,i,j) = (x_i, x_{i+1}, ..., x_{j-1}, x_j)$. $S_r$ was defined as $(\Pi(S_1), S_2, \Pi(S_3))$ and $S_f$ as $(ext(z,1,n), ext(z,n+1,2n), ext(z,2n+1,3n))$ where $z = \Pi(concat(S_1, concat(S_2, S_3)))$. The continuous line in Figure 3 shows the value of $I(S)$ as a function of the number of triples processed for the three sets. The dotted and dashed series in Figure 3 correspond to $I(S_f)$ and $I(S_r)$, respectively. The values of $I(S)$ of the control sets where higher than or close

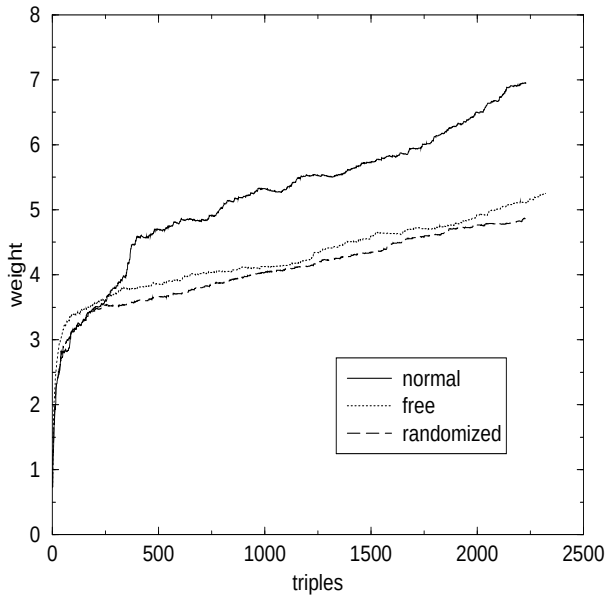


Figure 2: Total weight of the optimal trees as a function of number of triples processed.

to that of the original set. It seems that the learning technique performs better on $S_f$ and almost in the same way on $S_r$. Let us look into it.

We define $W(S)$ as

$$\sum_{i=1}^{n} weight\ of\ the\ heaviest\ tree\ of\ the\ i-th\ triple\ in\ S$$

The continuous line in Figure 2 shows the value of $W(S)$ as a function of the number of triples processed for the three sets. The dotted and dashed series in Figure 3 correspond to $W(S_f)$ and $W(S_r)$, respectively. Now, $W(S) > W(S_r), W(S_f)$ clearly. The calculation of the transinformation undergoes the participation of spurious links in the summation (our technique assumes that all possible links between members of the triple are possible), which are not taken into account when considering only the two heaviest links in the triple. Trees are more clearly captured in the normal set, as it should be.

Figure 4 plots the proportion of dominance of the three possible links of a triple $(x, y, z)$ as a function of the number of triples processed. It can be seen that the final amount of triples is enough for breaking the initial symmetry between the weights of the links.

Grammatical items can make the same word look different and rise the sparseness of data. Words could have been converted to canonic form before being placed in the triples. For example, by removing the '-s' ending of plural

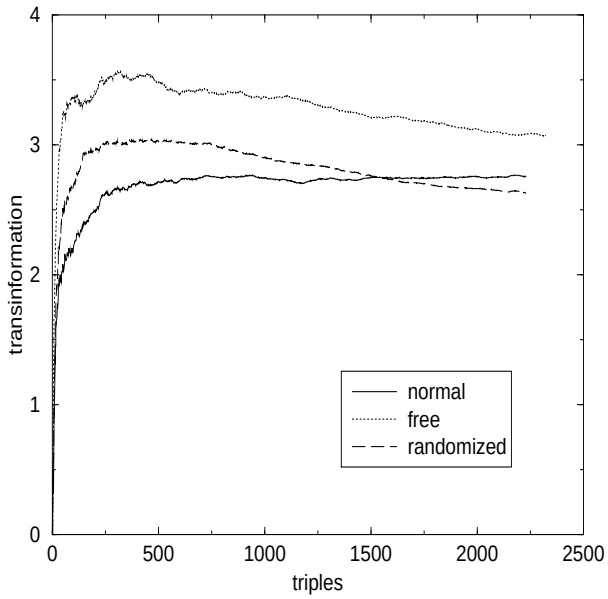


Figure 3: Transinformation as a function of number of triples processed.

nouns or the '-ed' termination of regular verbs. The transinformation of the set of triples with canonic words and that of the set with raw words happened to be very similar. We believe that the canonic form will be important for dealing with languages having plentiful grammatical items (e.g. Spanish).

# References

[1] Luis A. Nunes Amaral, Antonio Scala, Marc Barthélémy, and H. Eugene Stanley. Classes of behaviour of small-world networks. *Proc. Natl. Acad. Sci.*, 97(21), October 2000.

[2] Marc C. Baker. *Incorporation.* Chicago University Press, 1988.

[3] Chartrand. Cut-vertices and bridges. In *Introductory Graph Theory*, chapter 2.4, pages 35–49. Dover, New York, 1985.

[4] Noam Chomsky. *Lectures on Government and Binding.* Dordretch, Foris, 1981.

[5] Noam Chomsky. *The Minimalist Program.* MIT Press, 1995.

[6] W. Ebeling and T. Pöschel. Entropy and long-range correlations in literary english. *Europhysics Letters*, 26(4):241–246, May 1994.

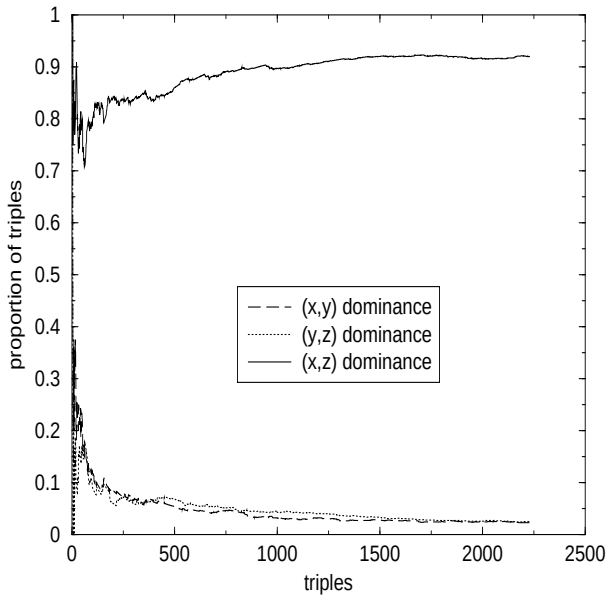


Figure 4: Proportion of triples where the links $(x, y)$, $(y, z)$ and $(x, z)$ are dominant as a function of number of triples processed.

[7] Joseph Emonds. *A Unified Theory of Syntactic Categories.* Dordretch, Foris, 1985.

[8] Charles Fillmore. *A Proposal Concerning English Prepositions.* Georgetown University Round Table on Languages and Linguistics, Washington D. C., Georgetown University Press, 1966.

[9] Norman M. Fraser. Prolegomena to a formal theory of dependency grammar. *UCKWPL*, 2:298–319, 1990.

[10] Teiji Furugori and Eduardo de Paiva Alves. Disambiaguation of syntactic structures using the strength of association in three word dependency relations. *Journal of Quantitative Linguistics*, 6:101–107, 1999.

[11] J. H. Greenberg. *Language Universals: with Special Reference to Feature Hierarchies.* Mouton, 1966.

[12] J. Gruber. *Lexical Structures in Syntax and Semantics.* Amsterdam, North-Holland, 1976.

[13] F. Harary. *Graph theory.* MA: Addison Wesley, 1994.

[14] Richard Hudson. *English Word Grammar.* Oxford, Blackwell, 1990.

[15] Ray Jackendof. *X-Syntax: A Study of Phrase Structure.* Linguistic Monograph, Cambridge, 1977.

[16] Ray Jackendoff. *The Architecture of Linguistic Faculty.* Cambridge, MIT Press, 1977.

[17] Richard Kayne. Datives in french and english. In *Connectedness and Binary Branching.* Dordretch, Foris, 1983.

[18] Hale & Keyser. On argument structure and the lexical expression of syntactic relations. In *The view from Building 20th.* Cambridge, Mass., MIT Press, 1993.

[19] Hilda Koopman. The structure of dutch pps. ms., 1993.

[20] W. Li and K. Kaneko. Long-range correlation and partial $1/f^{\alpha}$ spectrum in a noncoding dna sequence. *Europhysics Letters*, 17(7):655–660, January 1992.

[21] Wentian Li. Mutual information functions of natural language texts. *Santa Fe Institute Working Paper*, 1989.

[22] Igor Melčuck. *Dependencies Grammar: Theory and Practice.* New York, University of New York, 1989.

[23] Henk Van Riemsdijk. *A Case Study in Syntactic Markedness. The Binding Nature of Prepositional Phrases.* Dordrecht, Foris, 1982.

[24] Henk Van Riemsdijk. Categorial feature magnetism: the extension and distribution of projections. *The Journal of Comparative Germanic Linguistics*, 1(2):1–48, 1996.

[25] Duncan J. Watts and Steven H. Strogatz. Collective dynamics of 'small-world' networks. *Nature*, 393:440–442, June 1998.

[26] Deniz Yuret. *Discovery of Linguistic Relations Using Lexical Attraction.* PhD thesis, MIT, May 1998.

[27] Deniz Yuret. Lexical attraction models of language. Submitted to The Sixteenth National Conference on Artificial Intelligence (in http://www.ai.mit.edu/people/deniz/papers.html), 1999.

[28] G. K. Zipf. *Human behaviour and the principle of least effort. An introduction to human ecology.* Hafner reprint, New York, 1972. 1st edition: Cambridge, MA: Addison-Wesley, 1949.