

O TEOR DE INFORMAÇÃO SINTÁTICA: UMA MEDIDA DE MATURIDADE LINGÜÍSTICA PARA FALANTES DO PORTUGUÊS

José Marcelino Poersch
Professor de Lingüística
Aplicada da PUCRS

1 — APRESENTAÇÃO

A presente comunicação trata da investigação de uma medida, válida e operacional, de maturidade lingüística para falantes do português baseada na produção escrita de sujeitos discriminados em oito faixas diferentes de idade/escolaridade. Parte-se da hipótese de que o teor de informação sintática de um texto, baseada na freqüência de uso de certas estruturas bem como no tipo de relacionamento existente entre eles, pode constituir-se num bom índice de maturidade lingüística (razão dos determinantes). Objetiva-se validar, para o português, certas medidas de maturidade lingüística elaboradas com base em outras línguas bem como comparar estas medidas com a nova medida proposta.

Conclui-se que o índice **Razão dos determinantes** é um índice mais aceitável do que os anteriormente existentes. Calcula-se a equação da reta de regressão deste índice contra as faixas e constrói-se uma escala em três níveis (inferior, médio e superior) para cada faixa.

2 — PESQUISAS CORRELATAS

Muitas tentativas foram realizadas neste sentido. Duas forças contrárias parecem monopolizar e resumir as dificuldades encontradas pelos investigadores. De um lado objetiva-se encontrar os índices mais representativos, os de maior força discriminadora e em maior número possível; é o princípio da exaustividade. De outro lado existe a tendência de estabelecer as medidas com a maior simplicidade possível para facilitar a sua operacionalidade; é o princípio da simplicidade. O pesqui-

sador deverá determinar o ponto de equilíbrio entre estas duas tendências encontrando uma fórmula que seja, simultaneamente, válida e operacional. Muitos são os indicadores lingüísticos, principalmente no nível sintático, de maturidade lingüística levantados através de dados empíricos. O mais antigo destes indicadores é o **comprimento médio das frases** (11). Este indicador foi trabalhado, entre outros, por Stern (1907), Schlag (1917), Piaget (1923), Nice (1925), Davis (1937), Heider (1940). Hunt (1965) prova, através dos dados de sua pesquisa, que se trata de um índice pouco válido. Isso se deve à inexistência de uma conceituação precisa de frase e a possibilidade que uma criança tem de formular frases tão extensas quanto a própria redação, obliterando pontos e multiplicando as coordenadas sindéticas.

Lou LaBrant (1933), depois de ter verificado que o **comprimento médio de orações** (12), tanto principais quanto subordinadas, não se apresentava como uma medida significativa de desenvolvimento lingüístico, passou a analisar a relação entre orações subordinadas e o número total de oração; é a **razão de subordinação** (13). McCarthy (1954) também desenvolveu pesquisas nesta mesma linha.

Kellog Hunt (1965) retoma os índices anteriores e pesquisa a força discriminadora de cada um. Apresenta um novo índice: **comprimento médio de unidades termináveis** (14). Por unidade terminável entende uma seqüência de palavras que contém somente uma oração principal com todas as possíveis orações que lhe são ligadas. Verificou que este índice apresentava um coeficiente de correlação superior aos anteriores. Nesta pesquisa observa-se que, embora tenham sido emparelhadas as variáveis intervenientes de sexo e de nível intelectual, não é emprestado valor às variáveis idade e nível sócio-econômico-cultural. Por outro lado, também não deu importância ao tipo de dependências existentes entre os elementos constitutivos das unidades termináveis.

Com o advento do modelo gerativo-transformacional, outros indicadores foram investigados. A **complexidade transformacional** foi estudada por Miller (1962), por Menyuk (1963) e por Mehler and Bever (1974); este índice está relacionado com o número de transformações opcionais que a oração nuclear sofre para se superficializar.

A **teoria derivacional da complexidade**, estudada por Brown & Hanlon (1970) e, principalmente, por Fodor et Garret (1974), postula que a complexidade de uma frase é medida pelo número de regras gramaticais utilizadas em sua derivação.

A validade destas duas teorias é muito discutível por estarem baseadas na estrutura profunda, cujo status teórico foi questionado, e na oração nuclear, cujo número fica muito difícil de ser precisado em relação a uma determinada seqüência superficial. São "luas teorias constestáveis e sua força preditiva de resultados experimentais revelou-se medíocre" (Fodor et Garret, 1974). O modelo gerativo-transformacional constitui um modelo explicativo e não pode ter pretensões de ser um modelo psicológico de produção ou percepção da fala. Lemle (1977) diz que "apesar de alguns aparentes sucessos iniciais, a hipótese foi desconfirmada por experimentos posteriores, e os resultados dos experimentos iniciais podem ser reinterpretados em termos de mera estrutura superficial".

Alguns pesquisadores procuraram analisar a complexidade em termos de profundidade de encaixe (Yngue, 1960). Kimball (1973) levantou a hipótese da **complexidade ponderada**. Baseia-se na suposição que o processamento de constituintes da esquerda exige mais da memória imediata do que os constituintes da direita, uma vez que se trata de um registro cumulativo para a decodificação frasal. Com base nesta proposição, Kimball estabeleceu um cálculo de complexidade sintática, produtiva ou receptiva, que se processa linearmente da esquerda para a direita e de cima para baixo, na árvore que representa a estrutura superficial. O cálculo se efetua em unidades denominadas **própons**. Lemle (1977) aplicou este instrumento para a língua portuguesa. Nesta aplicação foram utilizadas frases previamente selecionadas contendo, de preferência, a ordem direta. A avaliação de sua adequação é feita intuitivamente, com base na competência do falante. Carece de dados empíricos que consigam emprestar um valor psicológico aos aspectos teórico-explicativos.

3 — NOVA PROPOSTA

O estatuto da comunicação lingüística analisa a mensagem como uma estrutura de signos verbais. Considera a mensagem como o resultado de um processo seqüencial que, simultaneamente, seleciona e insere. "Toda a estruturação da mensagem se baseia na possibilidade de escolha, ou seleção, a partir de uma série de alternativas" (Lyons, 1979 p. 71). As unidades de cada seqüência são selecionadas de um repertório (sistema lingüístico) de acordo com regras combinatórias e de acordo com as necessidades de comunicação ou de seqüências possíveis. Cada novo elemento a se encaixar na seqüência é o resultado de uma escolha entre as diversas possibilidades apresentadas pelos sistema.

Quanto maior o número de alternativas de escolha oferecido pelo sistema para determinado lugar da seqüência menos previsível será o elemento selecionado, menos provável será a sua ocorrência na mensagem; maior o seu teor de informação. Quanto mais um elemento puder ser previsto pelo sistema, menos informação sintática ele apresenta.

Segundo Zipf (1949), o sistema lingüístico é determinado por duas tendências: a tendência ao repouso e a tendência à eficácia. **A tendência ao repouso** (relacionado com o falante), leva a utilizar o mínimo de elementos diferentes e a dar a cada um o máximo de significação. **A tendência à eficácia** (relacionado com o ouvinte) exige, na mensagem, para o máximo de clareza e o mínimo de ambigüidade, a maior variedade possível de elementos, cada um com o mínimo de significado. Entre estas duas tendências estabelece-se um equilíbrio: lei do mínimo esforço, que consiste em informar o máximo com o mínimo de custo.

Pela tendência ao repouso estabelece-se uma relação inversa entre o comprimento de uma palavra e sua freqüência de uso. Palavras de poucas sílabas são mais freqüentes do que palavras de muitas sílabas. Pela mesma tendência, as construções simples, constituídas somente de elementos necessários, são mais freqüentes com que construções complexas, nas quais aparecem elementos opcionais.

Dentro do modelo estruturalista (Hjelmslev, 1975 e Martinet, 1971, mesmo utilizando terminologias distintas), convém notar que existem certas estruturas necessárias, por exemplo, os termos essenciais de uma oração (oração subordinada substantiva) e estruturas opcionais, como os elementos acessórios de uma oração (orações adjetivas e adverbiais, aqui denominadas de determinantes). Acredita-se que a utilização de estruturas necessárias independe da escolha do falante o que não acontece com as estruturas opcionais, supõe-se que a maior ou menor produção destas últimas estruturas constitua um indicador de maturidade lingüística.

Os elementos e as estruturas que necessariamente devem aparecer — por força do próprio sistema — são totalmente previsíveis, sua efetiva ocorrência, portanto, nada informa sintaticamente. Pode-se afirmar que o teor de informação sintática, presente numa mensagem, está diretamente relacionada com a quantidade de elementos acessórios que ela possui.

Por outro lado, as unidades mais freqüentes tem mais probabilidade de ocorrer na mensagem; sua efetiva ocorrência é

mais previsível e veicula menor teor de informação sintática do que as unidades menos freqüentes. Os polissílabos são menos freqüentes do que as palavras de menos sílabas. Sua ocorrência propicia maior informação sintática.

Em resumo, pode-se afirmar que a probabilidade de uma unidade ocorrer numa mensagem depende: 1º do número de alternativas que puderem preencher determinada função, determinado lugar; 2º da freqüência de ocorrência de cada unidade. Em relação ao primeiro, quanto maior o número de alternativas, menor a probabilidade para cada alternativa aparecer, maior a dúvida da escolha, maior surpresa o elemento escolhido produz, maior teor de informação veicula. No caso da interdependência entre duas unidades, o aparecimento de uma delas traz automaticamente a outra; o aparecimento desta última não constitui surpresa; não traz informação. Em relação ao segundo, quanto menos freqüente for o uso de determinados elementos, menos possibilidades tem de aparecer em certo lugar da mensagem, mais surpresa provoca o seu surgimento, mais informação veicula. O fato de que as palavras mais extensas são as menos freqüentes nos leva a supor que a ocorrência de polissílabos numa mensagem aumenta o seu teor de informação sintática.

As duas tendências presentes na formação de um sistema lingüístico (tendência ao repouso e tendência à eficácia) também estão presentes na aquisição da linguagem. A criança, à medida que aumenta o seu entrosamento social, é levada a utilizar cada vez mais as estruturas mais refinadas do adulto; este fato se baseia na necessidade que ela sente em se fazer entender sem ambigüidade e no nível do adulto.

A proposta da presente pesquisa é analisar a hipótese de que o teor de informação sintática, presente na produção escrita de um sujeito constitui um indicador válido de maturidade lingüística para falantes do português. A informação sintática será representada pelo número de polissílabos (POL) e das orações determinantes (ODET) em relação às unidades termináveis (orações plenas = OPLE): razão de determinação

$$(I_5) = \frac{\text{POL} + \text{ODET}}{\text{OPLE}}$$

4 — DEFINIÇÃO DO PROBLEMA

Pretende-se estabelecer uma medida de maturidade lingüística para falantes de português com sujeitos discriminados em oito faixas de idade/escolaridade (TABELA I), com base no

teor de informação sintática veiculada através de seu desempenho escrito.

TABELA I — DISCRIMINAÇÃO DAS FAIXAS SEGUNDO O BINÔMIO IDADE/ESCOLARIDADE

FAIXAS	VARIÁVEIS		
	IDADE	ESCOLARIDADE	
		1º GRAU	2º GRAU
F4	10	4a. série	
F5	11	5a. série	
F6	12	6a. série	
F7	13	7a. série	
F8	14	8a. série	
F9	15		1a. série
F10	16		2a. série
F11	17		3a. série

Maturidade lingüística, neste trabalho, é conceituada em termos de maior produção de determinados indicadores lingüísticos à medida que o indivíduo avança em faixas de idade/escolaridade. Estes indicadores podem ser encontrados nos diversos níveis de análise lingüística bem como nos níveis de articulação da mensagem.

Trata-se de estabelecer índices tais que sua regularidade de aparecimento em faixas diferentes seja capaz de medir o grau de maturidade lingüística de um sujeito.

Partindo das pesquisas correlatas bem como da proposta da presente investigação, são selecionadas cinco índices de maturidade lingüística (TABELA II) que constituirão objeto desta análise.

TABELA II — ÍNDICES DE MATURIDADE LINGÜÍSTICA

ÍNDICES	FÓRMULA	CONVENÇÃO
Comprimento médio das frases	PAL./FRASES	I ₁
comprimento médio das orações	PAL./ORAÇÕES	I ₂
razão de subordinação	SUB/ORAÇÕES	I ₃
comprimento médio das T-UNITS	PAL./T-UNITS	I ₄
razão de determinação	POL+O.DET/O.PLE	I ₅

As hipóteses de trabalhos são as seguintes:

1. — I₁, I₂, I₃ e I₄ correlacionam-se significativamente com as faixas de idade/escolaridade. Constituem índices válidos para medir a maturidade lingüística de falantes do português.

$$\left. \begin{array}{l} r_{11}^F \\ r_{12}^F \\ r_{13}^F \\ r_{14}^F \end{array} \right\} > 0,45$$

2. — I₅ correlaciona-se significativamente com as faixas de idade/escolaridade. Constitui um índice válido para medir a maturidade lingüística de falantes do português.

$$r_{15}^F > 0,45$$

3. — O coeficiente de correlação entre I₅ e as faixas de idade/escolaridade é significativamente maior do que os coeficientes de correlação referentes a outros índices. I₅ constitui um índice de maturidade lingüística mais significativa do que os demais.

$$rI_5^P > \begin{cases} rI_1^P \\ rI_2^P \\ rI_3^P \\ rI_4^P \end{cases}$$

A variável independente é constituída pelas oito faixas de idade/escolaridade. A variável independente compreende os diversos índices analisados. As variáveis intervenientes de sexo, de capacidade intelectual e de nível sócio-econômico-cultural foram devidamente emparelhadas, permitindo uma amostra satisfatoriamente homogeneizadas.

5 — AMOSTRAGEM E COLETA DE DADOS

Com a população da pesquisa foi definida em termos da variável idade/escolaridade, iniciando com a 4a. série e terminado com a 3a. do 2º grau, a amostra foi constituída pelos alunos de um estabelecimento do ensino de Porto Alegre com essa seriação completa. Ficou estabelecido que a amostra de cada faixa seria constituída por 20 alunos, dez de cada sexo. Para atender as exigências de homogeneização dos grupos e de outros aspectos imprevistos, o colégio escolhido teve que ter uma matrícula diurna mínima de 100 alunos por série, com uma distribuição bastante equitativa por sexo e com condições favoráveis para a aplicação dos diversos instrumentos. Entre os diversos estabelecimentos que preencheram todos os requisitos foi escolhido um. De cada aluno foram coletados os seguintes dados: número de chamada, série, data de nascimento, sexo, nível de inteligência e nível sócio-econômico-cultural e três redações. As redações foram escritas no segundo período do turno, como atividade letiva normal, sobre três assuntos: a escola, o esporte e a importância de ter amigos. Para o emparelhamento da amostra segundo as variáveis nível de inteligência e nível sócio-econômico-cultural decidiu-se utilizar somente os sujeitos classificados no nível médio, sendo eliminados os demais. Entre os sujeitos que preencheram todos os requisitos foram aproveitados, em cada faixa, os dez primeiros de cada sexo, seguindo rigorosamente a ordem de chamada. As redações destes cento e sessenta alunos constituíram o corpus da pesquisa. Deste corpus foram levantados os seguintes dados: 1. número de polissílabos, 2. número de palavras, 3. número de orações substantivas, 4. número de orações adjetivas, 5. número de orações adverbiais, 6. número de orações depen-

dentos, 7. número de orações dependentes determinantes, 8. número de unidades termináveis (orações plenas), 9. número total de orações e 10. número total de frases. A distribuição de cada item foi apresentada em tabelas, uma para cada faixa.

6 — ANÁLISE DOS RESULTADOS

A partir dos dados primários extraídos do corpus foram calculados os valores de cada índice para cada sujeito.

As fórmulas aplicadas são aquelas da TABELA II. Estes novos resultados foram tabulados por faixa e apresentados os diagramas de dispersão para cada índice: foram também calculados as médias dos índices de cada grupo e seu desvio padrão. A partir destes dados foram calculados as razões de correlação e os coeficientes de correlação (TABELA III)

TABELA III — MÉDIAS, DESVIOS PADRÃO, RAZÃO DE CORRELAÇÃO E COEFICIENTE DE CORRELAÇÃO DOS CINCO ÍNDICES

ÍNDICE	MÉDIAS	DESVIOS	RAZÃO DE	COEFICIENTES
	DOS ÍNDICES	PADRÃO	CORRELAÇÃO	DE CORRELAÇÃO
I ₁	19,481	6,936	0,3402 **	0,2735 **
I ₂	5,925	0,8153	0,2695 *	0,1911 *
I ₃	0,4529	0,1275	0,6044 **	0,5550 **
I ₄	11,464	3,141	0,6453 **	0,6000 **
I ₅	1,795	0,9116	0,7084 **	0,6771 **

Verificou-se que as correlações de I² apresentaram uma significância em nível inferior a 0,05 (5%) e que todos os outros índices apresentaram uma significância em nível inferior a 0,01 (1%). Estes dados permitem concluir que as correlações realmente existem na população e que não podem ser considerados como fruto do acaso.

O coeficiente mínimo de validade para um teste de usos práticos está ao redor de 0,45. Onde se conclui que um coeficiente acima de 0,45 é aceitável. No entanto, somente um coeficiente superior a 0,8 pode ser considerado forte. Com base nestas colocações podem ser analisadas as hipóteses 1 e 2. Os índices I¹ e I² apresentam-se como índices fracos para

medir a maturidade lingüística de falantes do português ao passo que os índices I³, I⁴ e I⁵ representam índices válidos, embora moderados. Nenhum deles pode ser considerado forte. O fato de eles não serem fortes não significa que eles não sejam importantes. Acontece que no levantamento destes índices de maturidade existem variáveis imponderáveis (motivação interna e externa, saúde, competência do professor, entre outros) que podem destruir uma alta correlação.

Para analisar a hipótese três faz-se mister comparar entre si os diversos coeficientes de correlação. Pela TABELA III verifica-se a existência de diferenças entre os diversos coeficientes de correlação e que o coeficiente de correlação relativo ao I⁵ é superior aos demais. O que importa, no entanto, é saber se esta diferença é realmente significativa. A comparação estatística fornece um valor de significância Z (TABELA IV).

TABELA IV — NIVEIS DE SIGNIFICÂNCIA DAS DIFERENÇAS ENTRE OS DIVERSOS COEFICIENTES DE CORRELAÇÃO

Coeficientes de correlação comparados	Valores de Z
r ₅ com r ₄	1,16
r ₅ com r ₃	1,75
r ₅ com r ₁	4,82

O valor crítico aceito para que uma diferença seja estatisticamente significativa está em torno de dois (2). Vê-se que a diferença de r₅ com r₄, bem como a diferença de r₅ com r₃, não é significativa. Somente a diferença r₅ e r₁ é significativa em nível inferior a 0,01 (1%). Desta maneira fica analisada a hipótese 3.

Resta agora estabelecer uma escala para a mediação da maturidade lingüística através de I⁵, razão de determinação.

O diagrama de dispersão configurado pelas variáveis I⁵ e F representa grosseliramente uma linha reta. Esta constatação visual necessita de um suporte matemático-estatístico.

As medidas de intensidade da relação entre duas variáveis analisadas são a razão de correlação e o coeficiente de correlação. Enquanto o coeficiente de correlação é uma medida de relação linear entre duas variáveis, a razão de correlação é uma medida tanto linear quanto não linear.

A razão de correlação sempre é maior do que o coeficiente de correlação. Se a discrepância entre estas duas medidas for muito grande, a melhor medida será a razão de correlação; no caso contrário, será o coeficiente de correlação. Normalmente aceita-se a discrepância de 2,92 como valor crítico de F no teste F. Os resultados encontram-se na tabela V.

TABELA V — VALORES DO TESTE DE F PARA OS CINCO ÍNDICES

ÍNDICES	VALORES
I ₁	1,17
I ₂	1,03
I ₃	2,11
I ₄	2,50
I ₅	2,29

Como não se registram discrepâncias estatisticamente significativas conclui-se que o coeficiente de correlação linear é uma medida tão boa da relação entre duas variáveis quanto a razão de correlação. Isto implica dizer que a equação desta relação é linear. Do cálculo da equação da reta de regressão I⁵ contra as F obtém-se a média para cada faixa (TABELA VI).

TABELA VI — VALORES MÉDIOS DOS INDICES PARA CADA FAIXA

FAIXAS	VALOR MÉDIO DOS INDICES
F4	0,85
F5	1,12
F6	1,39
F7	1,66
F8	1,93
F9	2,20
F10	2,47
F11	2,74

Para a construção da escala presume-se que os índices se distribuam normalmente dentro de cada faixa. Para cada F foram estabelecidos três níveis:

superior = 27%

médio = 46%

inferior = 27%

Como o grupo médio é constituído de 46% dos indivíduos, os índices do grupo médio se situam 0,4092 acima e abaixo do índice médio. Com base neste cálculo construiu-se a escala seguinte (TABELA VII):

TABELA VII: ESCALA DOS NÍVEIS DE MATURIDADE LINGUÍSTICA PARA CADA FAIXA

FAIXAS	NÍVEIS		
	INFERIOR	MÉDIO	SUPERIOR
F4	< 0,44	0,44 a 1,26	> 1,26
F5	< 0,71	0,71 a 1,53	> 1,53
F6	< 0,98	0,98 a 1,80	> 1,80
F7	< 1,25	1,25 a 2,07	> 2,07
F8	< 1,52	1,52 a 2,34	> 2,34
F9	< 1,79	1,79 a 2,61	> 2,61
F10	< 2,06	2,06 a 2,88	> 2,88
F11	< 2,33	2,33 a 3,15	> 3,15

7 — CONCLUSÃO

Da presente investigação conclui-se que a razão de subordinação (13), o comprimento médio das unidades termináveis (14) e a razão de determinação (15) são instrumentos válidos e suficientemente operacionais para medir a maturidade lingüística de falantes do português.

Conclui-se, outrossim, que a razão de determinação apresenta um coeficiente de correlação mais elevado do que os outros índices embora a diferença não seja estatisticamente significativa.

O fato de nenhum dos indicadores apresentar um forte coeficiente de correlação aponta para a necessidade de continuar as pesquisas no sentido de descobrir melhores índices.

Estes novos índices poderão ser descobertos na própria estruturação frasal: talvez na ordem das unidades utilizadas como, por exemplo, a intercalação dos advérbios entre o sujeito e o verbo bem como entre este e os objetos.

Poderão igualmente ser encontrados no teor de informação semântica veiculada pela mensagem. Será questão de encontrar um instrumento apropriado para realizar essa medição.

No entanto, parece que será na estruturação transfrasal, no nível da lingüística do texto, que poderão ser encontrados os melhores indicadores: relacionadores, elementos anafóricos, tipos de discurso, coerência interna, seqüencialidade lógica, entre outros.

REFERÊNCIAS BIBLIOGRÁFICAS

- 1 — BROWN, Roger & HANLON, Camile. "Derivational complexity in child speech". In: BROWN, Roger (ed) *Psycholinguistics*. London, Macmillan, 1970. p. 155-208.
- 2 — DAVIS, Edith A. "Mean sentence length compared with long and short sentences as a reliable measure of language development" *Child Development* (8): 69-79, 1937.
- 3 — FODOR, J. A. & GARRET, M. "Quelques déterminants syntaxiques de la complexité de la phrase". In: MEHLER, Jacques et NOIZET, Georges (eds). *Textes pour une psycholinguistique*. Paris, Mouton, 1974. p. 497-519.
- 4 — HJELMSLEV, Louis. *Prolegômenos a uma teoria da linguagem*. São Paulo, Perspectiva, 1975.
- 5 — HUNT, Kellog W. *Grammatical structures written at three grade levels*. Illinois, National Council of Teachers of English, 1965.
- 6 — KIMBALL, John. "Seven principles of surface structure parsing in natural language". *Cognition* (2): 15-48, 1973.
- 7 — LABRANT, Lou L. "A study of certain language developments of children in grades four to twelve, inclusive". *Genetic Psychology Monographs* (14) 387-491, nov. 1933.
- 8 — LEMLE, Miriam. "Complexidade sintática" In: LEMLE, Miriam & NARO, Anthony J. *Competências básicas do português*. Rio, Mobral, 1977. p. 51-148.
- 9 — LYONS, John. *Introdução à lingüística teórica*. São Paulo, Nacional, 1979.
- 10 — MARTINET, André. *Elementos de lingüística geral*. (5a. ed.) Lisboa, Sá da Costa, 1971.
- 11 — MCCARTHY, Dorothea. "Language development in children". In: CAR-MICHAEL, Leonard (ed). *Manual of child psychology* New York, John Wiley & Sons, 1954.

- 12 — MENYUK. "Syntactic structures in the language of children. *Child development*. (34) 407-422, march 1963.
- 13 — MILLER, George A. "Some psychological studies of grammar". *American Psychologist* (17,1): 748-62, jan. 1962.
- 14 — NICE, Margeret. "Length sentences as a criterion of a child's progress in speech". *Journal of Educational Psychology* (16) 370-379, 1925.
- 15 — PIAGET, Jean. *Le langage et la pensée chez l'enfant*. Neuchâtel, Delachaux, 1923.
- 16 — SCHLAG, J. "Grundsätzliches zu einen Häufigkeitwörterbuch der Kindersprache". *Ztsch, Päd. Psych.* (vol. XVIII): 216-226.
- 17 — STERN, C. *Die Kindersprache*. Leipsig, Barth, 1907.
- 18 — YNGUE, Victor H. "A model and an hypothesis for language structure". *Proceedings of the American Society*. (104): 444-466, 1960.
- 19 — ZIPF, K. *Behavior and the principle of least effort*. Cambridge (mass.), Addison Wesley, 1949.