

## Linguagem documentária na ótica de J.-C. Gardin



**Marilda Lopes Ginez de Lara**  
Universidade de São Paulo  
larama@usp.br

### 1 Introdução

Discursos recentes da Ciência da Computação falam em ‘acrescentar semântica’ à recuperação. Essa não é, entretanto, uma preocupação recente: Jean-Claude Gardin, há cerca de 50 anos atrás, buscava resolver problemas de um sistema ou codificação cujos termos necessariamente deveriam corresponder a uma definição suficientemente detalhada para que não subsistissem dúvidas quanto à sua respectiva significação (GARDIN, 1968). A esse sistema Gardin denominou ‘metalinguagem’, qualificando assim a linguagem usada para tratar e recuperar textos técnico-científicos.

O enfrentamento das questões da significação continua sendo um desafio, e revisitar o trabalho de Gardin pode não só permitir conhecer as bases teórico-metodológicas da representação do conhecimento por ele propostas ou sistematizadas, como realizar sua avaliação frente às iniciativas atuais. Necessário também ressaltar a importância de seu legado para o ensino: a produção de Gardin faz diferença na formação acadêmica e profissional em Ciência da Informação, porque assegura o entendimento dos fundamentos de uma atividade não raras vezes visto como meramente instrumental.

Em muitos aspectos, as reflexões de Gardin estavam à frente de sua época demonstrando, além do profundo conhecimento da literatura sobre a linguística e a lógica, pioneirismo na discussão e experimentação na área da semântica. Para a pesquisa de hoje, Gardin é importante porque toca em problemas ainda não

totalmente resolvidos e porque situa convenientemente a problemática da significação no universo da representação e da interpretação em linguagem.

Para organizar a exposição, após esta introdução localizaremos o termo 'linguagem documentária' frente a outra terminologia e discutiremos sua classificação como metalinguagem para destacar a ênfase que Gardin conferia à relação metalinguagem-*corpus* discursivo. Apresentamos, depois, uma tipologia dos léxicos, onde Gardin compara as ferramentas documentárias aos léxicos de apoio e, em seguida, os principais componentes da linguagem documentária, caracterizada por Gardin como metalinguagem. Procuraremos mostrar aspectos da discussão do autor sobre a análise semântica, destacando algumas de suas observações sobre a pertinência das propostas da linguística à época. Nas considerações finais destacaremos a atualidade de suas preocupações, apontando para a proximidade de algumas de suas propostas com as formuladas recentemente.

## **2 Linguagem documentária: denominação e classificação**

O termo linguagem documentária divide seu espaço com outros termos de significado aproximado – linguagem de indexação, vocabulário controlado etc. – cada um deles priorizando um aspecto diferente. Mais recentemente, a literatura em inglês registra o termo *Knowledge Organization Systems* – KOS, para agrupar uma série de recursos e instrumentos com funções diversas, mas também próximas: textos não estruturados (do tipo 'abstract'), listas de termos e conceitos (folksonomias, listas de autoridade, listas de *tags*, dicionários, glossários), listas estruturadas de conceitos (vocabulários controlados, sistemas de classificação, listas de cabeçalho de assunto, taxonomias, tesouros, ontologias) e, enfim, estruturas de relacionamento (mapas conceituais e modelos de referência de formas de organização) (SOUZA; TUDHOPE; ALMEIDA, 2012).

O W3C opta pelo termo genérico vocabulário, afirmando que na Web Semântica, os vocabulários definem conceitos e relacionamentos ou termos usados para descrever e representar uma área de interesse. Moreiro González (2011) também defende o uso da expressão vocabulários semânticos, justificando sua escolha pela atualização da terminologia no contexto da Web.

As variações devidas a ênfases diferentes, a ampliação da abrangência via KOS, ou a atualização específica da expressão linguagem documentária nesse âmbito, não apaga a diferença básica que marca o uso do termo 'linguagem' frente às outras denominações: a referência à linguagem natural, sistema complexo de comunicação cujas unidades têm seu valor definido relacionalmente no interior do sistema, ou seja, significam em função do conjunto (BENVENISTE, 1991). Vocabulários não são necessariamente estruturados: mais frequentemente são listas de ocorrências de um determinado *corpus* discursivo (GREIMAS; COURTÉS, 2008); já as linguagens documentárias o são, e sua melhor caracterização como tal depende da presença e da sofisticação de suas relações semânticas, bem como da potencialidade em gerar sintagmas via relações sintáticas. Não é sem razão que uma das definições mais precisas de linguagem documentária é a que a define a partir de seus elementos constituintes: um léxico, uma rede paradigmática e uma rede sintagmática (CROS; GARDIN; LEVY, 1968), elementos esses que respondem pela estrutura da ferramenta. O vocábulo 'documentária' (ou documental, em algumas versões), restringe, de fato, o termo linguagem documentária a um universo de documentos, podendo-se discutir sua flexibilização.

Na busca de um modelo de Análise Documentária, Gardin se refere a três fatores: "primeiro, à definição de análise documentária como extração de significado de um texto, o que implica uma referência a algo que é 'tirado' do texto e designado por símbolos que não são necessariamente os encontrados no texto – a saber, conceitos e suas interrelações, designados por símbolos *ad hoc* (descritores, indicadores de função, etc.). Para o sistema de símbolos usados para expressar o conteúdo, [o autor] propôs o nome de 'metalinguagem' " (GARDIN, 1969, citado por GARDIN, 1973, p.144). Em segundo lugar, à existência de similaridade entre as metalinguagens existentes à época, atribuindo suas variações às diferenças entre os campos e aos objetivos da representação. Gardin identifica, porém, analogias relativas: a) aos procedimentos seguidos para definir a unidade de análise, b) ao estabelecimento de relações analíticas entre os termos do conjunto, de modo a construir a organização semântica do léxico em questão, e c) à provisão de relações sintáticas entre os termos para conferir estrutura lógica à representação metalínguística das unidades sob análise. Em terceiro lugar, o autor se refere à distinção entre relação analítica (ou semântica) *versus* sintática (ou lógica). As relações semânticas são mais estáveis – como as que ocorrem em 'inseticidas',

diferentemente das relações ‘imprevisíveis ou sintáticas’, que são circunstanciais, acidentais, como em Drogas (Instrumento), Insetos (Paciente) (GARDIN, 1973). No entanto, a diferença entre as duas só tem fins práticos: “sob o plano sintático são veiculadas as relações novas, que são objeto do discurso, portanto não estabilizadas. Passa-se, no campo da ciência, progressivamente do plano sintático para o plano semântico: as relações novas, nesse sentido, ganham estabilidade e passam a constituir as relações semânticas” (GARDIN, 1973, p.144-145).

Tais fatores, em conjunto, justificam, para o autor, caracterizar a linguagem documentária como metalinguagem, porque ela põe em relevo o processo de construção e de transformação de textos que guardam contiguidade semântica entre si. Essa metalinguagem tem função normalizadora, operando semanticamente no tratamento de homonímias, homotaxias, homografias, polissemias etc. (GARDIN, 1970). Embora Gardin considere perigoso emprestar uma palavra que tem um sentido distinto na lógica, o termo metalinguagem é interessante, porque remete simultaneamente à linguagem de partida e à sua representação (em linguagem). Esses fatores justificam, também, sua posição relativamente à linguística chomskyana quanto à unidade de análise e quanto às suas contribuições para prover a linguagem com um esquema de articulações.

A necessidade de uma metalinguagem, para o autor, se deve às ‘anomalias’ da linguagem natural do ponto de vista semântico: sinonímia, homonímia, homografia, polissemias; alotaxias (variações de formulação das frases), equivalências mais complexas (entre a designação e sua definição, por exemplo). Os sistemas simbólicos estranhos à linguagem natural apresentam irregularidade de correspondências entre significantes e significados fundadas na invariância relativa dos significados nos domínios considerados (GARDIN, 1970, p. 632). “A metalinguagem ... não é outra coisa que esse sistema de símbolos, também chamado linguagem documentária, linguagem de recuperação especializada (*‘information retrieval language’*, em inglês)” (GARDIN, 1970, p.634).

### **3 Oposição entre os léxicos: lexicografia natural e lexicografia documentária**

Não menos importante para as reflexões sobre a organização da informação e do conhecimento é a proposta que posiciona as linguagens documentárias no conjunto dos ‘léxicos’. Isso demonstra que elas não foram vistas por Gardin como produtos totalmente separados da linguagem natural. A tipologia, anterior à proposta

de metalinguagem, já sinaliza a ênfase que o autor dará, no futuro, à relação entre linguagem natural e linguagem documentária e a especificidade das ferramentas de representação.

Os 'léxicos documentários' compreendem "todo conjunto de signos (palavras naturais, símbolos alfa-numéricos etc.) organizados ou não, que servem para construir representações indexadas de certos documentos" (GARDIN, 1966). Os produtos da 'lexicografia documentária' e os da 'lexicografia natural' (glossários e tesouros de língua), auxiliares da atividade documentária, são hoje reconhecidos como KOS.

Na construção das árvores que os organizam, a cada oposição, são discriminadas as diferenças entre os léxicos considerados. Os traços distintivos que permitem descrever as dicotomias são a correspondência com a linguagem natural (explícita ou implícita), a organização ou não dos léxicos (listas e classificações), a dimensão considerada (unidimensionais e pluridimensionais, unidimensionais reais e aparentes), o tipo de organização privilegiada (semântica, sintática ou mista). Sob mistas, as classificações são organizadas pela forma de organização (quase-hierárquicas, analítico-sintéticas e facetadas), sendo que as duas primeiras ainda são organizadas segundo a frequência de um termo na ferramenta (unívocas, multívocas).

As tipologias são sempre interessantes como meio para permitir visualizações, mas o mais importante dessa proposta de Gardin é que ela, ao recorrer a dicotomias num crescente processo de descrição semântica, reafirma o caráter estrutural da classificação e antecipa sua natureza metalingüística.

#### **4 Componentes da linguagem documentária**

Focalizando as linguagens documentárias em detalhe, Gardin identifica, no léxico, o componente mínimo da metalinguagem, conjunto de termos extraídos de uma dada linguagem natural (palavras-chave ou descritores). Um segundo componente é a estrutura paradigmática, conjunto de dados relacionados entre si *a priori* no léxico (GARDIN, 1965, citado por GARDIN, 1973). Esses dois componentes caracterizaram a maior parte das linguagens documentárias na sua configuração mais tradicional. Um terceiro elemento, uma estrutura sintagmática, foi adicionada por Gardin e equipe (CROS; GARDIN; LÉVY, 1968) na elaboração do SYNTOL – *Syntagmatic Organization Language*, para conferir maior flexibilidade e potencialidade

à ferramenta, bem como para dar conta da imensa diversidade de dados relacionados observada no processo de análise dos documentos (GARDIN, 1973).

O autor relativizou, no entanto, a distinção entre as duas estruturas (paradigmática e sintagmática) afirmando que ela era desejável para prover 'pontes' entre uma metalinguagem e outra quando da conversão intra-metalinguagem ou inter-metalinguagens (Aqui, uma antecipação da preocupação com a compatibilidade entre linguagens). A unidade relacional básica foi então definida, com base nas duas estruturas, como o sintagma mínimo 'Ri (x,y)', algoritmo de análise para formalização, onde x e y são termos do léxico e R é uma relação binária cujo significado é especificado por i. Regras de orientação são adicionadas para evitar ambiguidades.

Por meio desse sintagma mínimo é possível representar tanto relações paradigmáticas como sintagmáticas: no primeiro caso, relações estáveis entre os termos; no segundo, relações 'reais' que pressupõem interdependência entre dois significados. Tais relações são classificadas como dinâmicas (relações consecutivas ou de causalidade), estáticas (superposição associativa e predicativa entre conceitos), ou formais (coordenativas, comparativas e aproximativas).

O SYNTOL permite formular frases ao ligar duplas de palavras com base em suas relações mútuas de dependência. O sistema prevê a indicação do tipo de relação lógica existente entre as palavras consideradas (relação predicativa, associativa, consecutiva ou coordenativa), a sinalização do sentido da leitura do sintagma e a determinação das categorias formais respectivas (determinação da categoria num nível temático – lugar, tempo, modo – mais amplo do que o nível dos sintagmas, de modo a deixar claro o contexto). Operadores sintáticos são usados para evitar ambiguidade de interpretação, quando necessário, ou para ligar sintagmas próximos (CROS; GARDIN; LÉVY, 1968).

A proposta partiu, entre outras observações, da verificação de que as relações binárias são frequentemente utilizadas pelas metalinguagens para realizar a organização semântica, do mesmo modo que os indicadores de função (*role indicators*) o são para expressar relações sintáticas: os marcadores funcionais (predicados unários), cada um deles implicitamente referido a um ou vários marcadores semânticos numa dada representação metalinguística. Segundo Gardin, a vantagem do formato binário é que ele pode orientar cálculos lógicos para realizar deduções e inferências no processo de busca, quando especificações lexicais ou sintáticas devem ser gradualmente ampliadas para aumentar a revocação na

recuperação, ou ainda, na conversão de uma metalinguagem para outra, quando os mapeamentos (árvores sintagmáticas ou gráficos SYNTOL) das duas implicarem árvores de relações diferentes (GARDIN, 1973).

A experiência do SYNTOL, aqui apresentada de forma muito breve, mostra o avanço das pesquisas em Organização e Representação do Conhecimento numa época em que a maior parte das operações de sintagmatização era, no máximo, resolvida a partir do uso da lógica booleana.

## **5 A análise semântica**

Ao procurar por procedimentos destinados a colocar em evidência a significação de textos mais variados, Gardin (1970) mostra, já na década de 60 do século XX, a importância da análise semântica, quando os métodos de análise documentária eram reduzidos a operações empíricas de atribuição de palavras-chave a um texto. O autor discute os limites da linguística chomskyana – a teoria gerativo-transformacional – e se queixa de que raramente os linguistas falam de análise semântica.

Gardin critica, na teoria linguística dominante à época, a prioridade dada à análise sintática e à ‘estrutura de superfície’, que exerceriam papel prioritário na determinação do significado. Fala dos limites da estrutura profunda, questiona a sentença como unidade padrão de análise, o uso de categorias gramaticais (nomes, advérbios, adjetivos etc.) e de funções gramaticais (sujeito, objeto etc.) como meios para a expressar o significado. Cita outros autores (cujas contribuições foram publicadas entre 1968 e 1972) que também questionaram as propostas chomkyanas, a exemplo de Wilks e Hutchins.

Por outro lado, Gardin refere-se a uma série de experiências levadas a efeito por linguistas no mesmo período (início da década de 1970): a proposta de estruturas pré-lexicais (Postal), a adoção de ‘estratégias semânticas’ para guiar a interpretação de sentenças (Bever), a definição de ‘estruturas formacionais’ antes da análise dos enunciados (Grimes), uma unidade de análise menor do que a frase (Wilks) e, ao contrário, unidades mais amplas para prover uma melhor interpretação (sugeridas individualmente por Langendoen, Lakoff, e Thompson). O autor chama atenção para as pesquisas desenvolvidas na então URSS (Mel’chuck e Zholkovski, entre outros), que propuseram modelos de linguagem sofisticados com base em estruturas

semânticas e que, diferentemente de Chomsky que trabalhava do texto para o significado, iam em direção contrária, das estruturas conceituais para as expressões linguísticas (GARDIN, 1973).

Gardin se reporta também a outras experiências que considera interessantes, como a análise do discurso, a valorização da 'proposição lógica', os conceitos lógicos como conceitos que subjazem à fala, as redes semânticas (*semantic networks*) que incorporam informação relacional ou combinatorial 'sobre o mundo' (Lakoff; McCawley; Kuroda). Do mesmo modo, refere-se às 'pressuposições' da linguagem, ou regras de seleção que orientam o comportamento linguístico (citando Bever, Fillmore e Langendoen), as quais o autor situa em paralelo à noção de organização semântica ou paradigmática usada na análise documentária. Acrescente-se também a referência às categorias usadas na formulação de preposições (originalmente da gramática de casos, proposta por Fillmore, em 1968) reformuladas por Langendoen (com as '*roles structures*'), por Bever ('*basis relational structures*'), e por Wilks ('*mental templates*').

Toda a crítica feita à linguística gerativa transformacional foi feita em favor da lógica, não a lógica filosófica, mas a lógica simbólica, técnica. Ela forneceria o meio adequado de expressar a organização do pensamento, para categorizar os vocabulários (Agente, Paciente, Instrumento, Objetivo etc). Sobre as categorias, Gardin ressalta a prática documentária como antecipadora de muitas das propostas de uma semântica aplicada.

## **6 Considerações finais**

Como foi possível verificar, são várias as contribuições de Gardin para a compreensão e desenvolvimento das linguagens documentárias e para a representação de textos técnico-científicos. Seu trabalho, no entanto, não é isento de críticas: ao mesmo tempo em que ele era um homem à frente de seu tempo – enxergando na semântica os problemas para a veiculação da significação, enfatizando a importância da estruturação das metalinguagens etc., - devemos lembrar que ele foi inflexível quanto à abordagem dos discursos das ciências humanas de um modo geral, criticando-os por sua falta de objetividade e clareza a partir da referência da organização própria dos discursos das ciências naturais. Gardin enxergava como 'anomalias' a expressão da riqueza e complexidade do vocabulário



do discurso das ciências humanas, procurando meios para alcançar formas que retirassem o excesso de retórica que os caracterizava.

Inegável, no entanto, a contribuição do autor para o aprimoramento dos procedimentos da atividade com documentos, como se fosse necessário experimentar uma fase de rigor para chegar a revê-la, depois (como a afirmação de Wittgenstein, no *Tractatus*: “Deve, por assim dizer, jogar fora a escada depois de ter subido por ela” (WITTGENSTEIN, citado por GRAYLING, 1966, p.65).

Para finalizar, ressalto aqui a atualidade do projeto de Gardin quando se observa a similaridade de princípios que subjazem à proposta do sintagma ‘Ri (x,y)’, feita na década de 60 do século XX, e o RDF – *Resource Description Language*, linguagem para representação de metadados no formato de sentenças que relacionam entidades e propriedades. Não menos interessantes são suas preocupações sobre a compatibilidade e o mapeamento de metalinguagens, que correspondem hoje às iniciativas para promover a interoperabilidade entre vocabulários.

Ficaram fora desta exposição muitas contribuições não menos importantes e relacionadas, que seguiram as reflexões sobre as metalinguagens e as experiências sobre automação das análises. Uma delas é a proposta de esquematização de raciocínios ou ligações lógicas entre proposições descritivas (fatos e argumentação) e proposições interpretativas. Muito resumidamente, trata-se de um sistema que, a partir de uma base de conhecimentos, vai dos dados à interpretação através de um aparato lógico de cálculo, um motor de inferências, como um sistema especialista de inteligência artificial (GARDIN, 1987). Produto da análise logicista, é um dispositivo de maior formalização e sofisticação que permite passar, de uma indexação mais simples, à representação e à inferência (e interpretação). Mas esse é outro assunto.

## Referências

BENVENISTE, É.. *Problemas de linguística geral, I*. Campinas: Pontes; Ed. UNICAMP, 1991.

CROS, R.-C.; GARDIN, J.-C.; LÉVY, F. *L'automatisation des recherches documentaires: un modele général “Le SYNTOL”*. 2. ed. revue et augmentée. Paris: Gauthier-Villars, 1968.

GARDIN, J.-C. Elements d'un modèle pour la description des lexiques documentaires. *Bulletin des Bibliothèques de France*, 1966, p.171-182.

GARDIN, J.-C. Procédures d'analyse sémantique dans les sciences humaines. In: POUILLON, J. et MARANDA, P., orgs. *Échanges et communications: mélanges offerts à Claude Lévi-Strauss à l'occasion de son 60ème. anniversaire* (Tirage à part). Paris: Mouton, 1970, p.628-657.

GARDIN, J.-C. Document analysis and linguistic theory. *The Journal of Documentation*, v. 29, n.2, p.137-168, June 1973.

GARDIN, J.-C. Éléments d'un modèle pour la description des lexiques documentaires. *Bulletin des Bibliothèques de France*, v.11, n.5, p.171-182, 1966. Disponível em: <<http://bbf.enssib.fr/consulter/bbf-1966-05-0171-001>> Acesso em jun.2015.

GARDIN, J.-C. *Systèmes experts e publications savantes*. London: The British Library, 1987.

GRAYLING, A.C. *Wittgenstein*. São Paulo: Ed. Loyola, 1966.

GREIMAS, A.J. & COURTÉS, J. *Dicionário de semiótica*. São Paulo: Contexto, 2008.  
MOREIRO GONZÁLEZ, J.A. *Linguagens documentárias e vocabulários semânticos para a web: elementos conceituais*. Salvador: EDUFBA, 2011. 128p.

SOUZA, R.R.; TUDHOPE, D. ; ALMEIDA, M.B. Towards a taxonomy of KOS: dimensions for classifying Knowledge Organization Systems. *Knowledge Organization*, v.39, n.3, p.179-192, 2012.