Elementos de um modelo para a descrição de léxicos documentários

Marilda Lopes Ginez de Lara [I]
https://orcid.org/0000-0001-7888-9047

Nair Yumiko Kobashi [II]
https://orcid.org/0000-0002-3329-2535

[I] *Tradutora. Universidade de São Paulo, SP, Brasil.*
Professora Livre-Docente aposentada, ECA-USP.
Bolsista de Produtividade em Pesquisa - CNPq

[II] *Tradutora. Universidade de São Paulo, SP, Brasil.*
Professora Livre-Docente aposentada, ECA-USP.
Bolsista de Produtividade em Pesquisa - CNPq

http://dx.doi.org/10.1590/1981-5344/38979

Recebido em 29.03.2022 Aceito em 29.03.2022

Apresentação do autor

Jean-Claude Gardin (1925-2013) realizou pesquisas nas áreas da Arqueologia e da Documentação científica dos anos 1950 até seu falecimento em 2013. Na França, atuou como pesquisador do Centre National de Recherches Scientifiques (CNRS) e na École des Hautes Études em Sciences (EHESC) por mais de 50 anos.

Foi um pesquisador reconhecido não apenas na França, mas também na Europa Ocidental e Oriental, nos Estados Unidos da América, no Canadá e no Brasil, de um lado, pelas reflexões originais publicadas em livros e artigos, cursos ministrados e conferências proferidas em âmbito internacional e por sua atuação na concepção e elaboração de sistemas de recuperação de informação, tanto para a Arqueologia quanto para a Documentação.

Deve-se destacar que Jean-Claude Gardin tinha uma cultura vasta: formou-se em Economia na London School of Economics, após o término da II Guerra, tendo, em seguida, realizado estudos sobre as Ciências da Linguagem, na Sorbonne. Trabalhou com eminentes arqueólogos, historiadores, antropólogos, linguistas, lógicos e pessoa de Informática. Fernand Braudel e Claude Lévy-Strauss são, provavelmente os intelectuais mais conhecidos no espaço brasileiro.

Nesta apresentação serão destacados de forma breve, as suas contribuições ao campo da Ciência da Informação, notadamente os que se referem à "Análise documentária", que hoje integra o campo da Organização da Informação e do Conhecimento. Gardin foi um pioneiro do uso da informática nos processos documentários. Com efeito, trabalhou na Euratom e UNESCO, nas décadas de 1950 e 1960, na concepção e implementação de sistemas de automação de tratamento de documentos científicos. Os métodos da Inteligência Artificial, campo nascente do campo da Informática, à época, mereceram atenção especial de Gardin.

Suas contribuições são consideradas originais por ter proposto a associação da Semiologia e da Informática para lidar com a massa crescente de informação científica, fenômeno detectado no pós II Guerra Mundial. Citamos, a título de exemplo, três de seus trabalhos: a) o **SYNTOL** (Sintagmatic Organization Language) (GARDIN et al., 1964), que foi publicado na França e nos EUA, pela Rutgers University (EUA); b) **Linguistics and Documentation**, publicado no Journal of Documentation (GARDIN, 1973**)**, texto texto seminal no qual discute as possibilidades e limites da aplicação da linguística ao tratamento de textos/discursos para fins documentários, abordagem que contrastava com as práticas de tratamento estatístico de textos, vigentes à época; c) na obra Analyse des discours (GARDIN, 1974) Gardin aponta as especificidades da Análise documentária, comparando-a com outros tipos de análise, como a Análise literária, a Análise de conteúdo, a Análise logicista e a análise estrutural de contos.

Uma observação final: o texto "Éléments d'un modèle pour la description des lexiques documentaires" foi elaborado em 1965 para um grupo de trabalho envolvido em um projeto de indexação automática. Consideramos que o modelo de linguagem documentária de Jean-Claude Gardin é ainda atual. Esperamos que ele seja útil para melhor compreender os aspectos lógico-semânticos dos diferentes tipos de instrumentos de Organização e representação de informação.

Referências

GARDIN, J-C *et al*. *Le Syntol :* étude d'un système général de documentation automatique. Bruxelles: Presses académiques européennes, Bruxelles, 1964. 4v.

GARDIN, J-C. Document Analysis and linguistic theory. *Journal of documentation***,** v.29, n.2, p.137-168, 1973.

GARDIN, J-C. *Les analysis de discours*. Neuchâtel: Delachaux et Nestlé, 1974.

Título original: GARDIN, Jean-Claude. Elements d'un modèle pour la description des lexiques documentaires. *Bulletin des Bibliothèques de France*, p.171-182, 1966. https://bbf.enssib.fr/consulter/bbf-1966-05-0171-001.

1 Definições Preliminares

Recordemos, primeiramente, o que entendemos por *Léxico documentário***:** "todo conjunto de signos (palavras naturais, símbolos alfanuméricos etc.) organizados ou não, usados para construir representações indexadas de determinados documentos".

Os termos desta definição solicitam alguns esclarecimentos:

a) «signos»: em uma acepção deliberadamente geral, qualquer símbolo que designa os elementos de um conjunto lexical (palavras-chave, descritores, termos de indexação etc.), emprestados ou não de uma língua natural, codificados ou não, verbais ou numéricos etc.

b) «organizados»**:** apresentados em uma ordem inspirada no significado dos signos (ordem semântica, conceitual etc.). mas não em sua forma (ordem alfabética). Os léxicos «não organizados» são, portanto, os que não têm nenhuma estrutura semântica, sejam eles apresentados em uma ordem qualquer ou em ordem puramente formal (ordenação alfabética ou ordem numérica).

c) «representações» (indexadas): qualquer expressão das características dos documentos tratados: características de forma (tipo de publicação, formato, língua etc.) ou de conteúdo (disciplinas, assuntos, noções etc.) – por meio dos signos precedentes.

d) «documentos»: qualquer objeto, em sentido amplo (objeto concreto, imagem, texto etc.), considerado como unidade de análise e/ou de referência, em trabalhos de indexação.

Os léxicos documentários, assim definidos, apresentam uma grande variedade de formas e de denominações: glossários, tesauros, listas de palavras-chave, códigos semânticos, classificações etc. O objetivo desta observação é apresentar os elementos de um modelo que permitam destacar os traços distintivos de cada um, deixando de lado as analogias e

diferenças que possam ser observadas em um ou outro léxico, independentemente das variadas designações que são frequentemente dadas a eles.

Entre essas denominações, sem dúvida a mais usada, é <classificações>. Apresentaremos, inicialmente, uma tipologia geral dos léxicos documentários na qual as classificações são definidas como um grupo entre outros (§ II); em seguida, retomaremos esse grupo particular para tentar ordená-lo com base em um número restrito de propriedades estruturais (§ III).

2 Tipologia geral

A figura 1 resume os principais critérios da tipologia proposta:

1. Em primeiro lugar, a «lexicografia documentária» (isto é, o estudo dos léxicos documentários ou os seus modos de construção), se opõe à lexicografia natural, como a compreendem ou a praticam os linguistas em relação a qualquer língua natural. A primeira visa a estabelecer listas finitas de palavras1, sempre restritivas em relação ao vocabulário da língua natural das quais derivam, listas usadas para a formulação de representações indexadas, mais curtas e estereotipadas do que as expressões naturais a que correspondem. A lexicografia natural, por outro lado, não tem essa função redutora ou normalizadora: ela visa somente a observar, com base em um corpus particular, certos fatos que dizem respeito às ocorrências de palavras numa língua dada,

Consideramos necessário colocar a lexicografia natural nesta tipologia porque alguns de seus produtos são de fato utilizados em trabalhos documentários, ao lado dos léxicos documentários propriamente ditos; este é o caso particular dos glossários e dos tesauros definidos abaixo (§ 2 e 3).

2. Os «glossários» são definidos, por convenção, como quaisquer conjuntos de termos da linguagem natural apresentados em uma ordem não significativa, ordem alfabética, por exemplo: listas de termos técnicos, vocabulários especializados (bilíngues, multilíngues), micro glossários, dicionários científicos etc. Nesses glossários, os termos podem estar acompanhados de definições, de exemplos, de traduções etc.; por outro lado, toda indicação de equivalência ou de vizinhança entre os termos é em princípio excluída (sinônimos, formas canônicas etc.), na medida em que o glossário se aproximaria, então, de um «tesauro», num ou noutro dos sentidos abaixo (§ 3 e 5).

3. Esses mesmos conjuntos de termos naturais podem estar «organizados», isto é, agrupados em virtude de afinidades semânticas:

[1] Palavras" são aqui consideradas, num sentido vago, em lugar de "signos", para facilitar a apreensão das relações entre lexicografia documentária e lexicografia natural: referem-se às "entradas lexicais" que podem ser codificadas por meio de algum sistema de signos (N.A.).

dicionários de sinônimos, associações de ideias, dicionários conceituais, tesauros etc. É esta última palavra que retivemos para designar convencionalmente os léxicos naturais acompanhados de indicações semânticas desse gênero. Esclarecemos que a ordem de apresentação das entradas lexicais, nos tesauros assim definidos, pode ser tanto alfabética (ex.: a versão americana do Thesaurus de Roget), quanto conceitual.

4. A lexicografia documentária foi definida acima, em oposição à lexicografia natural (§1). Distinguimos duas grandes correntes: uma, a mais antiga, leva à constituição de listas de termos, organizados ou não, que devem servir à indexação documentária, sem que, no entanto, sejam explicitamente fornecidas as correspondências entre cada um desses termos e as palavras ou frases naturais usadas para representar. A outra, ao contrário, visa essencialmente a inventariar essas correspondências, de tal sorte que o processo de indexação pode ser padronizado para ser feito por qualquer processador automático. É este último gênero de inventário que definiremos em primeiro lugar, abaixo.

5. A partir dos léxicos naturais vistos anteriormente (§ 1 e 3), é fácil imaginar que se possa, na análise documentária, associar, a cada entrada, uma forma canônica destinada a servir de termo de indexação. Estabelece-se assim, um tipo de dicionário bilíngue (ou multilíngue), tendo «à esquerda», os termos ou expressões de uma ou de várias línguas naturais e, «à direita», as equivalências canônicas do léxico documentário adotado. Esses dicionários, que denominamos de «dicionários automáticos» no domínio da tradução automática, são frequentemente chamados de tesauros nos estudos consagrados à Documentação, por analogia aos dicionários conceituais vistos acima (§3), por serem os tesauros um desenvolvimento dos dicionários conceituais. Conservamos este uso propondo, contudo, distinguir por um sufixo as duas variedades de tesauros: LN, no primeiro caso (no qual somente a linguagem natural é considerada), LD, no segundo (que visa a substituir a terminologia natural de uma dada base lexicográfica por uma linguagem documentária).

6. A pesquisa sobre um sistema de correspondências inteiramente explícito entre termos naturais e termos de indexação é, contudo, um empreendimento muito recente não sendo possível hoje citar mais do que uma dezena de exemplos. Na maioria dos casos, os termos de indexação são apresentados sob a forma de listas independentes, sem fazer referência a uma língua natural: as correspondências permanecem implícitas tanto na mente dos autores quanto na dos usuários dessas listas. Esse é notadamente o caso das inúmeras «classificações» usadas em bibliotecas e centros de documentação, onde, em geral, o analista tem à sua disposição somente a sua própria cultura – "implícita" – para interpretar este ou aquele texto nos termos desta ou daquela classificação.

Uma maneira de objetivar a interpretação sem ter de construir os complexos dicionários de equivalência examinados no parágrafo precedente, é pela renúncia à «indexação» propriamente dita – em que frequentemente uma palavra ou uma expressão da linguagem natural é substituída por um outro termo [de uma linguagem documentária] – contentando-se com a mera «extração», em que são retidos um certo número de palavras ou frases do texto original, consideradas representativas de seu conteúdo, sem submetê-las a qualquer transformação2.

Tal seleção é sempre uma decisão que cabe ao analista (conforme o método dito do "lápis vermelho"…); além disso, ela é o resultado de certos cálculos estatísticos (cf. os métodos de H.P. Luhn e seus seguidores). Mas ela pode também ser obtida pela simples consulta a tabelas onde se encontram, em ordem alfabética, os termos permitidos (ou os excluídos) nas representações documentárias visadas. São essas tabelas que examinaremos agora.

7. Essas listas alfabéticas não devem ser confundidas com os "glossários" acima definidos (§.2): elas não se referem à totalidade dos termos naturais de um dado domínio científico ou técnico, sendo apenas uma seleção desses termos, retidos para a representação documentária. É neste sentido que elas pertencem à família dos "léxicos documentários", mais do que à dos "léxicos naturais".

8. O caso mais frequente é o das listas positivas, nas quais se enumeram – variantes morfológicas incluídas ou excluídas (cf. nota precedente) – as palavras e grupos de palavras consideradas, *a priori,* como dignas de serem retidas do enunciado natural de um documento (título, resumo ou texto integral), para representar o conteúdo. A análise deverá, portanto, basear-se na consulta a uma tabela, por um especialista ou por uma máquina.

As listas chamadas de "unitermos", nos Estados Unidos, são frequentemente definidas na perspectiva da extração. Pode-se, todavia, contestar a posição onde as colocamos, dentre os léxicos documentários, nos quais as correspondências são "implícitas": se a língua natural escolhida para designar os "unitermos" é a mesma dos textos analisados, essas correspondências não são de fato omitidas, a não ser que elas sejam obtidas por construção … Com efeito, são raras as listas que se enquadram em definições mais estritas:

[2] Ao menos no campo semântico, o único com que nos preocupamos aqui: por outro lado, é comum operar uma transformação morfológica para reduzir as diferentes formas de um mesmo termo, ou mesmo de mesma raiz, a uma base única.

a) as variantes morfológicas de bases registradas como "unitermos" são frequentemente implícitas;

b) da mesma forma, as equivalências na terminologia (ou na fraseologia) das diferentes línguas naturais são, contudo, submetidas a um processo de redução, tendo como referência essas mesmas listas.

Desse modo, uma grande parte deixada à interpretação do analista; por esse motivo, colocamos essas listas na categoria dos léxicos documentários sem correspondências explícitas com a linguagem natural, ao contrário dos tesauros precedentes (§5), distinguindo-os somente pela característica de apresentação "não-organizada", diferentemente das classificações (§. III).

9. Uma outra maneira de controlar a extração é estabelecer uma lista *a priori* de termos não permitidos: são então retidos, como critério para recuperação de um documento, todos os termos que não pertençam à tabela das palavras não permitidas (palavras-funcionais ou *stopwords*, enunciados indicadores de introdução ou clichês). Essas listas negativas são utilizadas principalmente para a fabricação automática dos índices de permutação, nomeados de diferentes formas (KWIC, Tabledex, Wadex, Physindex etc.).

N.B. Um dicionário "negativo" não define evidentemente um léxico documentário *stricto sensu*: este último é um "complemento" do dicionário, no sentido lógico do termo, em uma dada língua natural. Por outro lado, as correspondências entre léxico natural e léxico documentário não são necessariamente explícitas, pelas mesmas razões expostas acima (§8). É então por abuso, por assim dizer, que os dicionários negativos são aqui citados, considerando-se a analogia entre sua forma ("não-organizada") e o seu objetivo (a extração) em relação às listas precedentes (§8).

10. As listas "positivas" de palavras-chave organizadas em ordem alfabética são frequentemente acompanhadas de remissivas que lhes dão a aparência das classificações: escrever "a, ver também b, etc.", é uma maneira de indicar que há algum tipo de relação entre a e b para expressar uma ordem classificatória, tal como:

C (classe, termo genérico etc.)
- a
- b etc.,

ou ainda:

- a (termo genérico)
- b (termo específico) etc.

Essas mesmas listas são, por vezes, apenas um índice alfabético dos termos que pertencem a uma classificação determinada; examinaremos agora essas classificações.

3 Classificações

Entendemos "classificações" como todo conjunto organizado de termos destinados à indexação documentária, qualquer que seja o procedimento utilizado para expressar tal organização (remissivas, listas, codificações etc.).

a) "…termos destinados à indexação…: esta restrição é feita apenas para lembrar a oposição enunciada acima entre as organizações semânticas definidas num léxico natural ("tesauros, §3 acima) e as que são definidas dentro limites de um léxico documentário. O termo "classificação" é aqui reservado às segundas, por convenção.

b) "… qualquer que seja o procedimento utilizado para expressar a organização…": esta outra restrição visa descartar, ao menos num primeiro momento, as observações relativas aos códigos associados às classificações (sistema decimal, símbolos alfanuméricos, afixos, etc.).

Ë, com efeito, por seus traços estruturais, que parece razoável definir os principais tipos de classificação documentária, e não pelo uso que elas fazem de uma dada fórmula simbólica. Uma característica maior, sob este ponto de vista, é o número de relações analíticas[3] consideradas para organizar o léxico: é este o objeto da primeira distinção apresentada na fig. 2, entre classificações "unidimensionais" e "pluridimensionais".

1. Por "dimensões" de uma organização lexical entendemos a natureza das relações analíticas que a constituem. Uma classificação é unidimensional, por consequência, quando se leva em conta somente uma relação analítica. Este é o caso da imensa maioria das classificações feitas até hoje, em que apenas uma única relação é marcada pelo fato de um termo ser colocado em um dado grupo ou em uma classe, sem qualquer indicação explícita da natureza dessa relação, nem as diferentes interpretações de que ela pode se revestir de um grupo a outro.

De fato, a unicidade da "dimensão" é, com frequência, apenas aparente; sob a qualificação "hierárquica" dada a essa relação única, descobre-se, sem dificuldade, diferentes tipos de relações – qualificação, localização, instrumentalidade etc. – que nada têm em comum com a

[3] Por "relação analítica" entendemos a relação que une um termo à classe de que faz parte numa organização lexical, mesmo quando essa relação não é definida de maneira explícita.

relação de inclusão. É por esta razão que propomos estabelecer uma distinção entre unidimensionalidade "real" ou "aparente":

2. A unidimensionalidade real, por exemplo, está presente em sistemas taxonômicos, tais como os construídos nas ciências naturais, onde a relação constitutiva é efetivamente única, em todos os níveis e para todos os grupos da classificação. Pelo que conhecemos, as classificações em uso na documentação não são unidimensionais dessa maneira, a não ser localmente, para certas partes da organização (ex..: nomenclaturas de espécies naturais, de elementos químicos etc.); além disso, existem outros tipos de relações apenas implícitas que permanecem anônimas, confundidas, por essa razão, com a relação hierárquica estrita.

3. O anonimato das relações analíticas mascara, com efeito, sua diversidade, nas classificações onde a unidimensionalidade é só aparente, como ocorre com frequência. Um esforço de análise será particularmente desejável para resgatar os diferentes tipos de relações implicitamente consideradas na arquitetura de cada forma de organização; é, com efeito, na escolha dessas relações (seu número, sua natureza, sua função) que se manifesta a originalidade de uma classificação do ponto de vista estrutural e, consequentemente, seu lugar na tipologia dos léxicos documentários.

4. Uma vez separadas as distintas categorias de relações analíticas subjacentes a uma classificação, ela aparece como uma organização pluridimensional, em que cada termo pode ser associado a várias classes, em virtude das diferentes "dimensões". Atualmente, o exemplo mais desenvolvido é, sem dúvida, o "Semantic Code" da Western Reserve University (W.R.U), dos Estados Unidos.

Uma vez definida a pluridimensionalidade *a priori* (como no Semantic Code) ou, ao contrário, estabelecida *a posteriori* (como se sugere acima), falta encontrar um princípio de compartilhamento entre os inúmeros e diferentes tipos de classificações pluridimensionais. O que propomos fundamenta-se na distinção empírica entre duas categorias de relações analíticas: de acordo com seu fundamento (ou sua função na linguagem documentária), de ordem "semântica", ou de ordem "sintática". Definamos agora o sentido que damos a esses termos.

5. Por ordem semântica, entendemos uma ordem dos termos (ou seja, dos objetos ou noções que eles designam) que reflita um conjunto de definições correntemente admitidas por um grupo humano (por exemplo, a comunidade científica do séc. XX etc.). Se se admite que essas definições visam a exprimir certas características inerentes das entidades consideradas, ou seja, sempre válidas (ou pertinentes) e, qualquer que seja o contexto no qual se as considere, a classificação adotada pode ser concebida como a imagem de uma ordem (provisoriamente, localmente) natural, onde a significação primitiva dos termos determina seu lugar na

organização geral. É neste sentido que as denominamos "semântica", em oposição a um outro tipo de organização, dita "sintática", que definiremos a seguir.

6. Por organização sintática entendemos uma ordem dos termos fundada não na essência das entidades que eles designam – que se manifesta através de um corpo de *definições* – mas sobre sua função particular num campo determinado de observação (ex.: uma classificação do "iodo" entre os "materiais utilizados na fabricação de produtos farmacêuticos" etc.)[4]. As propriedades consideradas para agrupar os termos são contingentes, no sentido de que são válidas somente em um contexto restrito, ao qual a classificação tenta se adaptar (no exemplo acima, "a fabricação de produtos farmacêuticos").

Esse é o caso das organizações ditas "facetadas", ao menos no sentido restrito que se deve dar ao termo para manter alguma especificidade em relação à noção geral de classe ou de categoria semântica[5].

7. Se é fácil conceber organizações exclusivamente "semânticas" ou "sintáticas", segundo os critérios precedentes, de fato, a maioria das classificações em uso na documentação parecem ser mistas: os pontos de vista "essenciais" e "funcionais" são tanto alternados quanto combinados, em proporções e ordem, dentro das variadas ordens sequenciais. Sem dúvida, é a este gênero de classificações mistas que se aplica a denominação "analítico-sintéticas" usada em certos manuais; a separação das duas categorias de relações é, em todo caso, um meio cômodo de ordenar tais classificações numa escala contínua, desde a ordem estritamente "semântica" das taxonomias mais puras, até a ordem exclusivamente sintática de certos léxicos facetados.

8. Neste último caso – classificações facetadas – qualquer termo pode, de fato, aparecer em várias regiões do léxico, de acordo com as diferentes classes funcionais às quais se pretende ligá-las (ex.: o iodo como "matéria-prima", como "produto final", como "reagente" etc., em um léxico relativo à indústria farmacêutica). As mesmas repetições são observadas nas organizações "semânticas", tal como as definimos acima, e substancialmente pelas mesmas razões. Assim, dentro de um sistema taxonômico qualquer, a recorrência de certos grupos de termos (por exemplo, uma seção "Membros" repetida em diferentes capítulos A, B, C

[4] Especificamos que se trata da ordem classificatória, apesar do emprego da palavra "sintática" e não de relações lógicas observadas entre os termos de indexação na representação de documentos particulares (papéis, ligações etc.) (N.A.)

[5] Sobre esse assunto, ver: J.- C.Gardin. Free classifications and faceted classifications. in Classifications research, Proceedings of the second international conference (FID/CR Committee on classification research), Elsinore 14-18 sept. 1964, ed. by P. Atherton. - Copenhague, Musksgarrd, 1965, pp.161-168 (N.A.).

... consagrados às espécies animais etc.) serve para marcar, de fato, uma relação sintática (membros de A, membros de B etc.), de uma outra ordem que não as relações hierárquicas da taxonomia. A repetição dos mesmos termos em diferentes classes de um léxico é, então, um índice de que este último é organizado de maneira a dar conta das relações não somente semânticas, mas de certos tipos de relações gramaticais, como os léxicos facetados. Pode ser útil, em consequência, distinguir entre as classificações ditas unívocas**,** nas quais um termo ocupa um lugar e apenas um, e as classificações multívocas, com repetições. Esse é o sentido da última dicotomia proposta na parte de baixo da figura 2.

9. Outras características são também concebíveis, relativas, por exemplo, aos próprios termos de indexação dessas classificações:

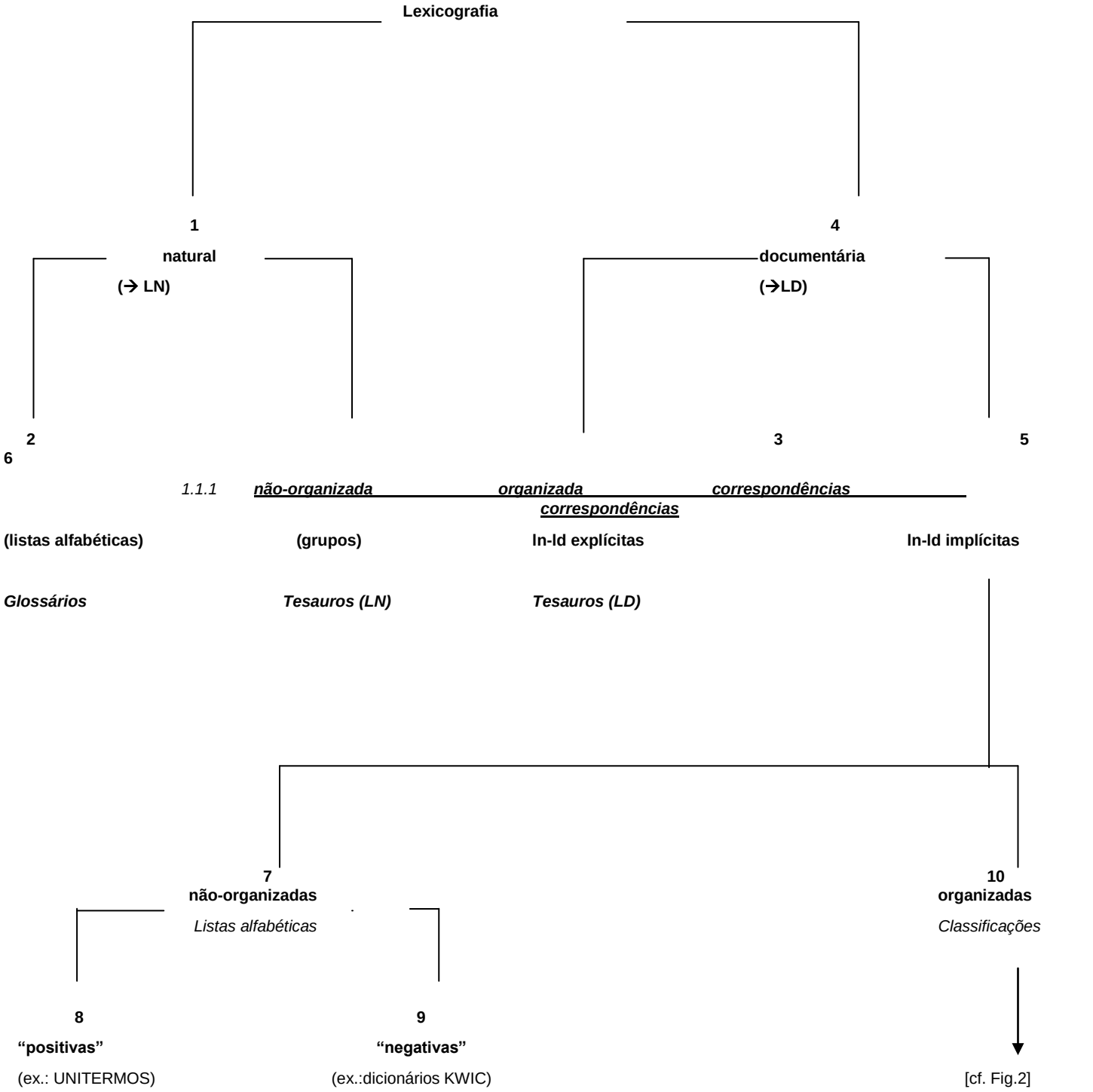
- grau de generalidade semântica;
- função lógica eventual (ex.: "descritores compostos");
- forma codificada etc.

ou sobre o aspecto quantitativo do léxico considerado globalmente (número de termos, número de classes, número médio de níveis de inclusão etc.). Parece, contudo, que a pesquisa das propriedades estruturais definidas acima seja mais fecunda para a descrição fina dos léxicos documentários.

Jean-Claude Gardin,

Diretor da Section d'automatique documentaire du

Centre National de la Recherche Scientifique.



LN – Linguagem natural

LD – Linguagem documentária

FIGURA 1

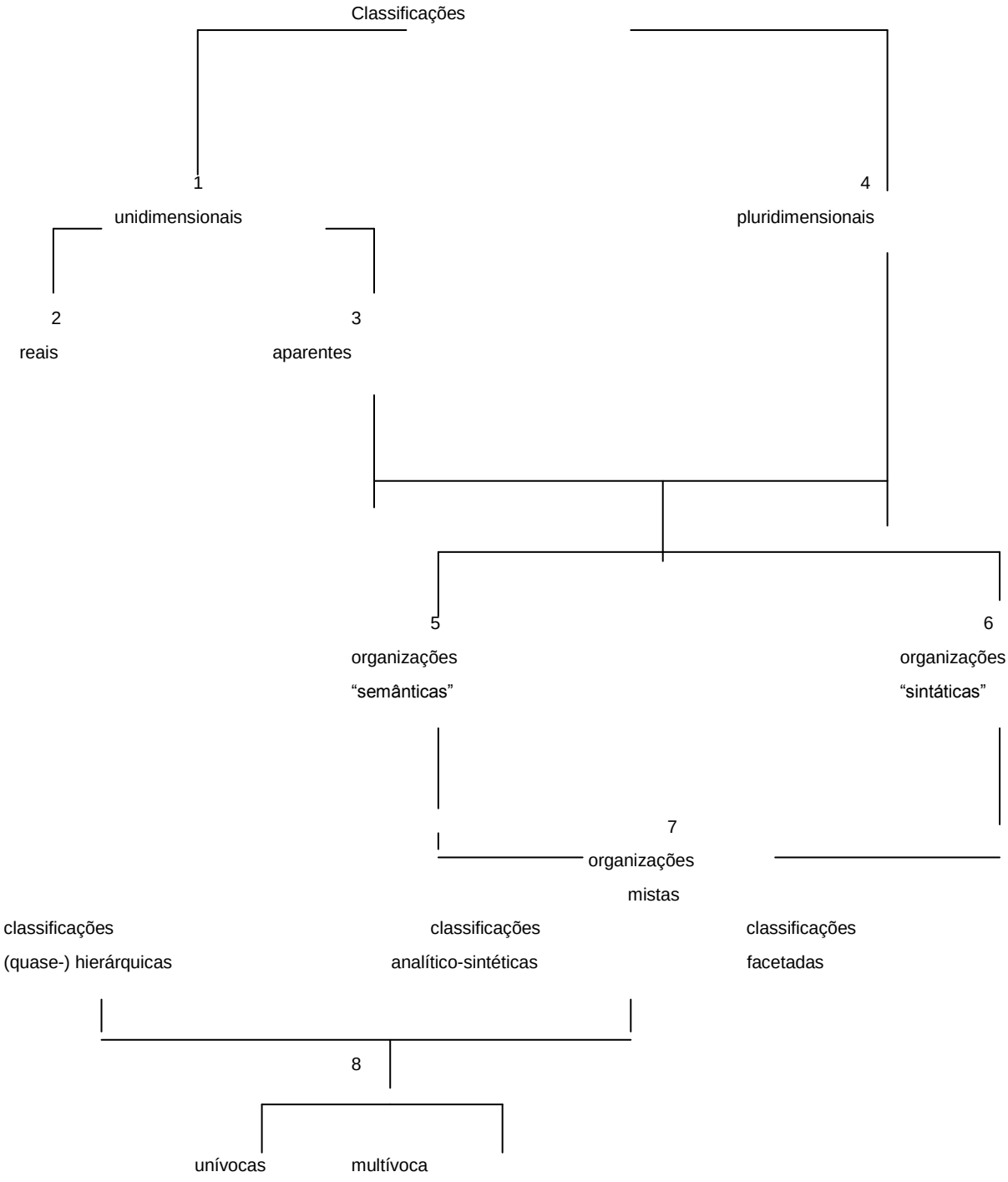


FIGURA 2

Agradecimento

Agradecemos à Profa. Cristina Dotta Ortega pelos esforços para obter a autorização para a tradução do artigo junto aos responsáveis pela obra de Jean-Claude Gardin, na França.