Problèmes linguistiques et informatiques liés à l'établissement d'un thésaurus juridique

par

Luigi PARENTI
Daniela TISCORNIA
CNR - FIRENZE

La responsabilité de cette communication est partagée entre Mlle Daniela Tiscornia et moi-même.

L'objet de cette communication est la présentation d'un "thesaurus" expérimental sur le droit pénal, conçu et élaboré par une équipe de chercheurs au sein de l'Istituto per la Documentazione Giuridica del Consiglio Nazionale delle Ricerche (Institut pour la Documentation Juridique du Conseil National des Recherches), qui a son siège à Florence.

Afin d'épargner la patience de ceux qui nous écoutent, en considérant que tous les présents connaissent la fonction instrumentale d'un "thesaurus" dans la recherche des données dans une banque de données documentaires, nous ferons seulement une brève allusion aux différences entre les "thesaurus" dits "a priori" et les "thesaurus" dits "a posteriori". Nous avons élaboré notre "thesaurus" en fonction d'une banque de données bibliographiques de doctrine juridique, recueillies par notre Institut. Notre banque de données se distingue des autres du même genre parce qu'elle fournit à l'usager non seulement les données bibliographiques relatives aux articles de doctrine juridique qui paraissent périodiquement dans la presse, spécialisée ou autre -c'est-à-dire la presse d'opinion, journaux, revues de sociologie, de politologie, de politique du droit etc.-, mais aussi un "abstract", bref mais exhaustif, du sujet traité. Le système STAIRS de l'IBM est utilisé pour interroger notre banque. Le support informatique de notre "thesaurus" est le système VS/TLS qui, comme vous savez, est une amplification et une amélioration du système de recherche STAIRS.

Nous vous parlerons d'abord des problèmes rencontrés sur le plan linguistique, en approfondissant surtout l'aspect morphologique et, partiellement, l'aspect sémantique; nous illustrerons après, brièvement, les caractéristiques du programme TLS.

Par "thesaurus" nous entendons un lexique organisé du langage documentaire. Mais qu'est-ce le langage documentaire dans le cadre du droit ? Par langage documentaire nous entendons un langage artificiel, c'est-à-dire un langage par mots-clés, utilisé pour l'établissement de l'index et la recherche des données, obtenu par la sélection du langage juridique naturel; par langage juridique naturel nous entendons le langage spécifique du droit, auquel il faut ajouter la partie du langage naturel qui est utilisée en droit, parfois en maintenant les mêmes valeurs du langage commun, parfois en revêtant un sens particulier, exclusivement technique.

Le "thesaurus" peut être établi "a priori", en fonction de l'index établi des documents à mémoriser. Il sera composé d'un index de termes et d'expressions dont la valeur sémantique sera définie de telle façon que leur emploi ultérieur permettra de distinguer un champ conceptuel spécifique, soit pendant la phase de l'établissement de l'index, soit pendant la phase de l'interrogation. Quand le "thesaurus" est établi "a posteriori", il s'insère dans une banque de données déjà existante. Il est donc plus difficile d'arriver à dominer le langage, étant donné que dans des archives documentaires comme les nôtres, n'importe quel terme mémorisé peut être une clé de recherche. Nous avons établi deux critères-guide pour la sélection des descripteurs :

1) élimination des paroles grammaticales (articles, prépositions), des noms propres et de ces termes techniques non-ambigus qui, à cause de leur extrême fréquence, alourdiraient la liste inutilement (ex. : delitto, pena, etc.);

2) adjonction au "thesaurus" de tous les termes technico-juridiques, outre les termes appartenant au langage naturel qui, en examinant les contextes, assument une importance juridique aux fins de la recherche. Afin d'expliquer la définition du "thesaurus" comme un lexique organisé, il faut savoir que, même dans les cas d'un "thesaurus a posteriori", la question doit être établie de telle façon qu'elle exprime un concept défini. Pour cette raison, nous avons établi des liaisons linguistiques et conceptuelles à travers une structure réticulaire avec des relations horizontales et verticales.

RELATIONS LEXICALES

Nous avons appelé "équivalence syntactique" la relation qui existe entre des termes de la même famille lexicale (verbe, substantif, adverbe), là où la superposition sémantique est totale; la relation d'"équivalence syntactique" sert aussi à lier un terme avec ses variantes graphiques et avec ses flexions grammaticales (en effet, le TLS prévoit, outre le "thesaurus" même, le Linguistic Integrated System, c'est-à-dire un analyseur linguistique qui permet l'extension automatique de la question aux formes -déclinaisons et conjugaisons-; malheureusement, il n'existe pas pour la langue italienne : quand il sera fait, l'usage de descripteurs normalisés sera possible). Mais nous avons séparé des équivalents syntactiques les termes qui, tout en appartenant à une famille lexicale commune, ont une gradation sémantique assez accentuée, telle qu'elle génère une expansion trop diffuse de la recherche (ex. : adozione, adottare/adottabilità), avec la relative formation de "bruit". Nous avons lié ces termes à leurs "parents" morphologiques avec une relation (CFR) d'avertissement à l'usager qui veut étendre sa recherche.

RELATIONS SEMANTIQUES

Comme nous avons déjà dit, le langage du "thesaurus" est un langage artificiel, c'est-à-dire élaboré à des finalités documentaires et les regroupements sémantiques sont créés "a posteriori" sur la base de l'usage des termes dans les documents aux archives : ceci explique les difficultés que nous avons rencontrées à cause des phénomènes de la polysémie, de l'homographie et de la synonymie. La polysémie et l'homographie risquent de produire du bruit dans la recherche documentaire. Souvent dans le langage juridique, le polysens naît de l'adoption des termes du langage commun qui sont utilisés avec un sens technique très précis (ex. : azione). Dans un "thesaurus a posteriori", il n'est pas possible de trouver des solutions rigoureuses, telle l'exclusion ou du moins la limitation la plus stricte de l'usage de polysens et d'homographies, l'adoption de signes graphiques de distinction des sens et ainsi de suite. Il nous a semblé que la seule solution possible était celle de signaler à l'usager la multiplicité des valeurs sémantiques, en ajoutant des codes à chaque valeur du descripteur homographe ou polysens.

Nous définissons comme synonymes deux termes interchangeables, c'est-à-dire tels que, si on remplace l'un par l'autre dans une phrase, la valeur sémantique de la phrase reste inchangée. Dans le langage juridique, les synonymes sont très rares, puisque les termes techniques du droit, qui expriment des concepts juridiques très précis, sont définis. Il y a pourtant des termes du langage naturel utilisés dans un sens technique (particulièrement lorsqu'il s'agit de doctrine), qui peuvent produire des phénomènes de synonymie (ex. : assassinio SYN omicidio), et des termes qui sont parfois utilisés improprement comme synonymes (ex. : locazione SYN affitto). Le "thesaurus", par sa nature instrument de recherche de données, ne peut pas ne pas tenir compte de ces liaisons. Nous avons donc pensé qu'il était indispensable de distinguer deux gradations d'affinité sémantique : une, plus restreinte, comprenant les véritables synonymes, en y incluant même les termes qui ne sont pas italiens, mais qui ont une correspondance exacte dans notre langue (ex. : termes latins souvent utilisés dans la pratique juridique, comme *bonorum cessio, exceptio rei iudicatae*, etc.), l'autre, plus large, comprenant les termes presque synonymes ou voisins.

Une relation ultérieure que nous avons distinguée est celle que nous pouvons appeller la relation de "hiérarchie sémantique". C'est-à-dire, nous avons mis en rapport entre eux, des termes génériques et des termes spécifiques (ex. : reato - delitto; delitto - furto; furto - furto aggravato) afin de permettre à l'usager d'effectuer une recherche plus large et plus articulée.

Il faudrait parler aussi de l'opposition sémantique (relations d'antonymie, inversion, complémentarité) et de la corrélativité (ou association d'idées). Nous nous bornerons à observer que la relation hiérarchique et celle de la corrélation peuvent être établies à un niveau plus conceptuel qu'étroitement sémantique, à tel point de pouvoir être définies relations systématiques dans ces cas. Mais nous préférons laisser le soin d'un traitement plus approfondi de ces sujets à notre collègue, M. Taddei-Elmi.

Nous illustrerons brièvement les caractéristiques du support informatique du "thesaurus". Il est à souligner que le "thesaurus" existe indépendamment du dictionnaire STAIRS, même si, évidemment, celui-ci contient les descripteurs en tant qu'extraits des documents.

Le programme offre la possibilité d'utiliser, pour lier les descripteurs, des relations prédéfinies, ou d'en créér d'autres, en en définissant les propriétés selon les besoins de l'usager. Une relation peut être définie symétrique, transitive, converse (ex. : dans les relations d'hiérarchie : A *Narrower term* B, B *Broader term* A). Il en suit une génération automatique de descripteurs-mots d'entrée. Parmi les relations de système, nous ferons particulièrement allusion à la relation de *Multiple word*, applicable à des descripteurs composés de deux ou de plusieurs termes (ex. : mandato di cattura) : elle produit la scission automatique des termes du syntagme (les mots vides étant exclus), qui deviennent donc des nouveaux descripteurs. La relation de synonymie SYN entraîne la liaison du champ sémantique entier d'un descripteur à son synonyme. Cette propriété, basée sur la présomption de la synonymie parfaite, peut être cause d'erreur dans le cadre du langage juridique. Nous avons donc choisi le remplacement du SYN de système par la relation EQU d'équivalence sémantique, en limitant les

propriétés et les implications de celle-ci.

Pendant la phase de recherche, le système TLS offre plusieurs possibilités à l'usager :

1) examiner le "thesaurus" et donc effectuer la recherche textuelle avec les procédés STAIRS normaux;
2) sélectionner dans le "thesaurus" les descripteurs utiles et donc élaborer la question en se servant des opérateurs booléens;
3) rappeller les relations du "thesaurus" dans l'interrogation directe des archives, en étendant automatiquement la question. Dans tous les trois cas, il est possible de compléter la question, en activant le Linguistic Integrated System, et d'étendre automatiquement la recherche à toutes les formes -déclinaisons et conjugaisons- des descripteurs (lorsqu'on disposera, pour l'italien, du programme IBM).