LIGIA CAFÉ
IBICT/CNPq

UMA PROPOSTA FUNCIONALISTA PARA A ANÁLISE DE UNIDADES TERMINOLÓGICAS COMPLEXAS (UTC)

Résumé

Notre recherhce vise à contribuer à la conception d'un logiciel d'extraction des unités terminologiques complexes (UTC) contenues dans des textes en langue portugaise (variété brésilienne). Pour ce faire, nous avons choisi comme corpus de données les UTC contenues dans des textes écrits en portugais sur la biotechnologie des végétaux. Ces UTC sont décrites et analysées selon leurs aspects morphologiques, syntaxiques, sémantiques et pragmatiques à la lumière de la théorie de la grammaire fonctionnelle de Simon Dik.

Palavras-chave: *unidade terminológica complexa; gramática funcional; extração automática de termos.*

Delimitação da pesquisa

A presente pesquisa trata da descrição e da análise das unidades terminológicas complexas em língua portuguesa (variante brasileira). A idéia de desenvolver uma pesquisa desta natureza teve como origem o projeto de criação do Banco de dados terminológico do Brasil (Brasilterm). Analisando os diversos estágios do funcionamento de um banco terminológico, concentramos nosso interesse na etapa do tratamento automático do corpus escrito com o objetivo de extrair segmentos terminológicos. Uma vez que o estudo do conteúdo para fins de automação é uma área bastante vasta, limitamos nossa pesquisa à análise do comportamento linguístico dos diversos componentes que formam um tipo específico de segmento: as unidades terminológicas complexas (UTC).

O corpus desta pesquisa é formado de UTC de base nominal encontradas no português escrito do Brasil. Os textos (manuais, artigos científicos, relatórios de pesquisa e dissertações) utilizados como fonte de coleta pertencem à área dos processos relativos à Biotecnologia de Cultura de Tecidos de Plantas.

Fundamento teórico

Esta pesquisa utiliza como fundamento teórico a Gramática Funcional de Simon Dik (1978, 1980, 1981, 1983,1987, 1989). Esta teoria postula três princípios essenciais para uma análise coerente de uma língua, quais sejam: a concepção da língua como instrumento de interação social, o reconhecimento do papel primordial da pragmática na análise da língua e o estudo da sintaxe fundado na semântica. Do ponto de vista metodológico, adotamos, no âmbito deste estudo, o modelo para análise das expressões linguísticas apresentado por Dik. Este modelo se baseia na interpretação do segmento por meio da análise da predicação. Este método resulta em regras de formação que sistematizam os dados linguísticos sobre os segmentos terminológicos, no nosso caso as UTC, de forma bastante eficaz para a compreensão do funcionamento das expressões no seio da língua.

A teoria da Gramática Funcional de Dik foi concebida para a análise da língua geral, mais precisamente para estudar a estrutura interna das frases e suas funções semânticas, sintáticas e pragmáticas. Uma vez que nosso estudo é de caráter terminológico, tivemos que adaptar a regra de formação e as três funções apresentadas por Dik, sem, no entanto, nos afastar dos principais fundamentos desta teoria.

Regra de formação

Para a Gramática funcional, a língua geral é formada de predicados e de termos (argumentos e satélites), o que constitui a estrutura geral da predicação de uma língua. No que concerne as línguas de especialidade, as UTC são definidas como segmentos formados de uma *base* seguida de *argumentos* et/ou *satélites*. A *base* é o centro da estrutura com o qual os *argumentos* tem uma relação direta. Como ilustrado na figura 1, este conjunto composto de *base* + *argumento* forma uma *predicação nuclear*. Os *satélites*, por sua vez, são elementos que têm por função completar o significado produzido da relação entre a *base* e o *argumento*. Desta

forma, os *satélites* mantêm essencialmente uma relação com o conjunto *base + argumento*, ou seja com a *predicação nuclear*. A união de um *satélite* à uma *predicação nuclear* é chamada *predicação estendida*.

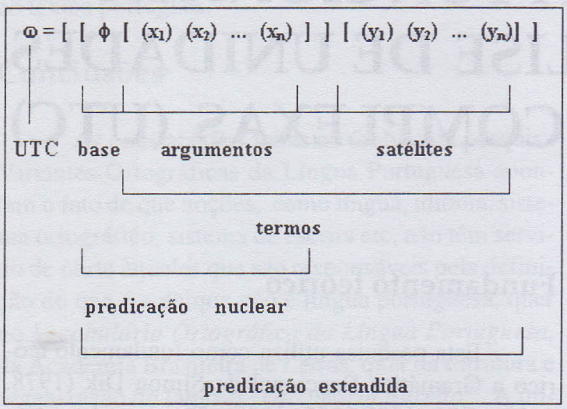

***Figura 1** - Regra de formação das UTC*

Os índices 1, 2 e *n* indicam a ordem dos *argumentos* e dos *satélites* na UTC. Os colchetes são utlizados para circundar os conjuntos e sub-conjuntos de relações estabelecidads entre os elementos da UTC. Estas relações podem existir no interior da *predicação nuclear* ou da *predicação estendida*. Pode-se igualmente encontrar relações dentro de um conjunto de argumentos. Neste último caso, utilisamos a barra oblíqua (/) para monstrar a hierarquia relacional entre os argumentos. Este signo gráfico é utilizado principalmente para marcar a expansão de um *argumento*, ou seja, no caso em que um *argumento* é modificado ou especificado por um outro *argumento*. A determinação de uma regra de formação para uma UTC é fundamentada num conjunto de interpretações de cunho funcionalista. A figura 2 ilustra a análise da predicação da UTC *propagação clonal in vitro*.

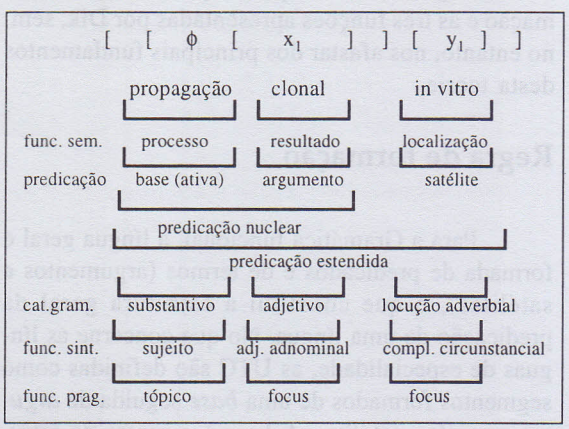

***Figura 2** - Exemplo de análise funcional de UTC*

A primeira análise é semântica, isto é, identificamos a função semântica de cada componente da UTC. Assim, *propagação* representa o papel de *processo*, *clonal* o papel de *resultado* e *in vitro* de *localização*. Esta interpretação é fundamentada no conhecimento de que propagação é um processo que resulta em clones e é feita in vitro (em laboratório). A partir destes dados, podemos aplicar a análise da predicação e assim definir que *propagação* é a base do segmento pois é à partir dela que podemos definir o papel semântico dos outros componentes da UTC. Além disso, a base é dita ativa pois se trata de um substantivo derivado de um verbo.

A direita de *propagação* encontra-se o elemento *clonal*. Este componente é considerado um argumento por dois motivos:

a) mantém uma relação direta com a base,
b) é um elemento fundamental exigido pela semântica da predicação. Sem ele, a informação seria incompleta.

O elemento seguinte a ser analisado é *in vitro*. Este é classificado como satélite pois mantém uma relação com todo o conjunto de elementos à esquerda do segmento (*propagação + clonal*), isto é com a predicação nuclear como um todo. O conjunto da predicação nuclear mais o satélite forma a predicação estendida.

Esta interpretação nos leva a determinar a regra de formação desta UTC, qual seja [[ø [x_1]] [y_1]].

Prosseguindo a análise, determinamos para cada componente da UTC a sua categoria gramatical, função sintática e função pragmática. No que se refere as categorias gramaticais, a UTC *propagação clonal in vitro* é formada de Subst. + Adj. + Loc. Adv. Quanto as funções sintáticas, *propagação* representa o sujeito da UTC. Esta função é atribuída a todos os elementos de uma UTC que servem de ponto de partida para a análise do segmento sob a prespectiva sintática. O argumento *clonal* preenche a função de adjunto adnominal. Esta interpretação do argumento é baseada em dois critérios :

a) se relaciona com uma base ativa,
b) possui categoria gramatical adjetiva.

O satélite *in vitro* é um complemento circunstancial. Reservamos esta função a todos os componentes de uma UTC cuja categoria é uma locução adverbial.

No que se trata das funções pragmáticas, *propagação* é interpretada como o tópico pois é o elemento ao qual todos os argumentos e satélites fazem referência a fim de precisar o objetivo de comunicação. *Clonal* e *in vitro* são ditos focus pois são elementos que especificam o estatuto informacional do elemento tópico, ou seja, ele pontua a informação do tópico.

Resultados

A análise do comportamento linguístico do corpus resultou em 210 UTC representadas por 10

regras de formação. No âmbito deste trabalho, concentraremos nossa exposição nas principais características de cinco regras de formação mais representativas do corpus da pesquisa.

A análise das UTC formadas pela regra [ϕ [x]] revelou que a grande maioria das *bases* apresentadas neste tipo de formação são estáticas, isto é, *bases* que não são derivadas de verbos. Um exemplo de base estática é o componente *genoma* na UTC *genoma nuclear*.

Observamos igualmente que esta regra se apresenta principalmente sob o padrão [N [SA]]. Uma das características interessantes deste padrão é o de regrupar as UTC formadas pelo prefixo de negação *não*, como é o caso da UTC *bactéria não virulenta*. O corpus conta também com UTC compostas segundo o modelo [N [SP]]. Esta fórmula sintática comporta algumas UTC que possuem um artigo no interior de sua estrutura morfológica (ex. *axila da folha*). Finalmente, gostaríamos de ressaltar o papel da preposição contida na estrutura [N [SP]] que tem o valor de marcar as relações de significado dentro da predicação nuclear.

Os diversos componentes da regra [ϕ [x_1]] possuem diferentes papéis sintáticos. Assim, a *base* constitui o sujeito da UTC enquanto que os *argumentos* podem ocupar tanto o lugar de adjunto adnominal como de complemento nominal ou de complemento circunstancial. A UTC *base genética*, por exemplo, tem argumento de função sintática adjunto adnominal enquanto que a UTC *mapa de restrição* possui argumento com papel de complemento nominal. Já o argumento de *propagação em massa* preenche a função de complemento circunstancial.

Ainda no que concerne a regra [ϕ [x_1]], a análise de nossos dados resultou na identificação de treze tipos de relações conceituais fruto do relacionamento estabelecido entre as diversas funções semânticas das bases e dos argumentos das UTC. No que diz respeito ao emprego destas funções semânticas, a função *entidade* é principalmente utilizada pelas bases (41%) enquanto que os argumentos empregam esta função em apenas 5% das UTC analisadas nesta regra.

Quanto as regras [ϕ [y_1]] et [[ϕ [x_1]] [y_1]], é a presença de satélites que as distingue dos outros modos de formação. Os satélites caracterizam uma *predicação estendida*, ou seja, uma predicação que matém uma relação com a *predicação nuclear* como um todo e não apenas com a base como é o caso dos argumentos. Observamos, no entanto, que somente a regra [ϕ [y_1] representa uma relação direta do satélite com a base devido a um provável apagamento de um argumento no curso do processo de formação da UTC. Este seja talvez a explicação para a UTC *cultura in vitro*.

Ao nível gramatical, os satélites podem ser sintagmas preposicionais ou locuções adverbiais. No nosso corpus, estas locuções adverbiais são representadas por expressões latinas do tipo: *in vivo*, *in vitro* e *in situ*. No plano sintático, os satélites se caracterizam como complementos circunstanciais e, ao nível funcional, eles são classificados como locuções adverbiais estendidas. Esta última etiqueta reflete a localização deste elemento dentro da estrutura da predicação. Do ponto de vista conceitual, os satélites podem exercer dois tipos de funções semânticas : *localização* ou *estado*. Eles se relacionam com os outros componentes da UTC de forma a representar determinadas relações conceituais.

A presença da categoria gramatical locução adverbial resultou em novos tipos de modelos morfossintáticos para as regras [[ϕ [x_1]] [y_1]] e [ϕ [y_1]]. Além da formula [[N [SP]] [SP], a regra [[ϕ [x_1]] [y_1]] apresenta os padrões [[N [SA]] [SAdv.]] (*propagação clonal in vitro*) e [[N [SP]] [SAdv.]] (*cultura de anteras in vitro*). No que diz respeito a regra [ϕ [y_1]], somente a fórmula [N [SAdv.]] (*polinização in vitro*) caracteriza o padrão morfossintático deste modo de formação.

As regras [ϕ [x_1 [$x_{1/1}$]]] e [ϕ [x_1 [$x_{1/1}$ [$x_{1/1/1}$]]]] têm como característica principal a expansão do primeiro argumento. No caso da regra ([ϕ [x_1 [$x_{1/1}$]]], o argumento é modificado apenas uma vez (ver fig. 3), enquanto que no caso da regra [ϕ [x_1 [$x_{1/1}$ [$x_{1/1/1}$]]]], tanto o argumento é modificado quanto a sua expansão (ver fig. 4).

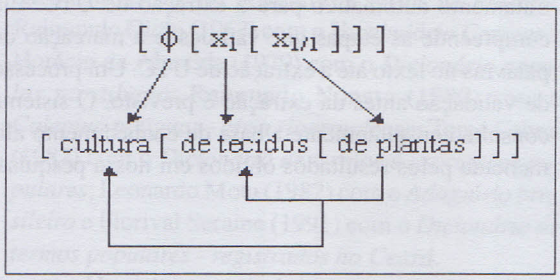

Figura 3 *- Exemplo de UTC para a regra* *[ϕ [x_1 [$x_{1/1}$]]]*

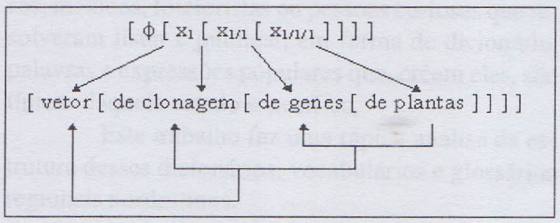

Figura 4 *- Exemplo de UTC para a regra* *[ϕ [x_1 [$x_{1/1}$ [$x_{1/1/1}$]]]]*

Este fenômeno é interpretado em nosso estudo como um processo de encadeamento de UTC. Na verdade, se trata de uma inserção de uma UTC dentro da estrutura de uma UTC principal. É um recurso utilizado para formar um determinado conceito à partir de conceitos e de estruturas terminológicas jà existentes.

A presença de argumentos-advérbios (*planta **genotipicamente** uniforme*) e de expansões-advérbios (*família de seqüências **moderadamente** repetitivas*) foi verificada nestes dois tipos de modos

de formação. Estes advérbios são formados de um adjetivo mais o sufixo *-mente*. No plano funcional, eles são interpretados como advérbios nucleares pois ocupam um espaço na predicação nuclear. No caso específico da regra [ϕ [x_1 [$x_{1/1}$]]], estes advérbios podem ser formados de um adjetivo proveniente da língua de especialidade ou da língua geral. No que se refere as relações conceituais, a relação semântica do advérbio é estabelecida com o adjetivo que o segue. É justamente a sequência advérbio + adjetivo que especifica ou modifica a base da UTC.

Conclusão

Nossa pesquisa toma como base, com vimos, uma metodologia de cunho funcionalista. Com base na Gramática funcional de Simon Dik, testamos nosso método em uma amostragem de termos da área da Biotecnologia de Plantas.

A partir das adaptações feitas ao modelo de Dik e nos resultados obtidos, propomos em nossa pesquisa um meio ambiente informático fundamentado em Inteligência Artificial e orientado para a modelagem de conhecimentos linguísticos[1]. Esta proposta é esquematizada sob a forma de uma estrutura geral do tratamento automático para a extração de UTC. Ela compreende as etapas que vão desde a marcação de palavras no texto até a extração de UTC. Um processo de validação antes da extração é previsto. O sistema consulta constantemente a base de conhecimento alimentada pelos resultados obtidos em nossa pesquisa.

Referências bibliográficas

CAFÉ, Ligia, 1999, *La description et l'analyse des unités terminologiques complexes en langue portugaise (variété brésilienne) : une contribution à l'automatisation de la Banque de données terminologiques du Brésil (Brasilterm)*, Québec (Canadá), Université Laval, Tomo I e II. (Tese de doutorado).

DIK, Simon, 1978, *Functional Grammar*, North-Holland, 230 p. (North-Holland Linguistics Series, 37).

_____, 1980, *Studies in Functional Grammar*, London, Academic Press, 245 p.

_____, 1981, Predication and Expression: the Problem and the Theorical Framework, Dans : *Predication and Expression in Functional Grammar*, London, Academic Press, p. 1 - 17.

_____, (éd.), 1983, *Advances in Funcional Grammar*, Foris Publications, 415 p. (Publications in Languages Sciences, 11).

_____, 1987, Some Principles of Functional Grammar, Dans : *Funcionalism in linguistics*, Amsterdam / Philadelphia, John Benjamins, p.81 - 100. (Linguistics & Literary Studies in Eastern Europe, 20).

_____, 1989, *The Theory of Functional Grammar*, part I: The Structure of the Clause, Dordrecht - Holland / Provindence RI - USA, Foris Publications, 433 p. (Functional Grammar, 9).

[1] Esta proposta pode ser encontrada no corpo da tese de Café (1999) referenciada na bibliografia.