Dicionauro: Dicionário Infantil Multilingue e Multimédia

Noélia Maria F. dos Santos e Manuel António R. Rodrigues
Instituto de Letras e Ciências Humanas
Departamento de Informática
Universidade do Minho

Abstract

The aim of this study is to establish a basis for a children's dictionary, including definitions in Portuguese and translations in English, French, Spanish and Dutch. It was needed an extended research and review of open source resources to build a terminological database named as *macroestructure*, in which was implemented a *microestructure* held by relevant semantic relations, either to the childlike public or to the informatics approach. The combination of multilingual and multimedia resources with computational lexicography will allow the existence of a dynamic electronic dictionary, flexible to the users' needs.

Keywords: lexicography, open source resources, terminological database, ontologies, information extraction
Palavras-chave: lexicografia, recursos *open source*, base de dados terminológica, ontologias, extracção de informação

1. Introdução

Nos últimos anos têm sido feitos avanços significativos na área de Processamento de Linguagem Natural (PLN), o que tem permitido a criação de ferramenta informática variada, desde processadores de textos, dicionários em suporte electrónico, tradutores *online*, entre outras. Associar a lexicografia à nova ferramenta informática é cada vez mais uma prática incontornável pois facilita e agiliza em larga escala etapas importantes no fazer lexicográfico, tais como: a extracção, o armazenamento e o tratamento de grandes quantidades de informação. No domínio da lexicografia é inevitável falar-se de grandes quantidades de informação e, como tal, existe a necessidade de poder geri-las de modo prático e eficaz. Se a informação a ser tratada pelo lexicógrafo estiver organizada numa base electrónica, isso representa uma grande vantagem na medida em que, a qualquer momento, o lexicógrafo poderá editar ou reorganizar o material de várias formas. Por exemplo, se um dia quiser extrair uma lista de todos os verbos existentes na base, através de um programa ou *script* poderá facilmente conseguir isso, o que representa uma rapidez e eficiência máximas em comparação com o tempo que levaria a fazer essa extracção manualmente. Para que isto seja possível, cada item

***Textos Seleccionados. XXIV Encontro Nacional da Associação Portuguesa de Linguística*, Lisboa, APL, 2009, pp.483-495**

deverá estar "codificado" ou numerado, para que possa, mais tarde, ser identificado e extraído. Isto implica algum exercício de antecipação das necessidades que podem surgir no futuro.

1.1. Objectivos

O objectivo principal deste projecto é o de criar um dicionário infantil multilingue (Português, Inglês, Francês, Espanhol e Alemão) sustentado por um método que alia todas as potencialidades das novas tecnologias às exigências que a elaboração de um dicionário infantil representa. Essas exigências prendem-se com a necessidade de fornecer à criança uma ferramenta útil e divertida, um dicionário que apresente definições simples, mas explicativas, tendo em vista as suas necessidades de aprendizagem.

A elaboração de um dicionário infantil requer um cuidado especial de elaboração e "cortar" uma definição que servia a um público em geral não pode constituir um método lexicográfico adequado na construção de definições destinadas ao público infantil, pois essa tendência de redução, em vez de simplificar, apenas priva a criança de um conjunto de informações úteis à apreensão e compreensão de conceitos. Pensamos que podemos aproveitar todas as potencialidades que nos são fornecidas pelas novas tecnologias para recolher, tratar e organizar a informação disponibilizada e colocá-la ao serviço, tanto do público infantil, como, numa posição mais abrangente, do público em geral e estamos convictos que uma política de disponibilização livre de conteúdos só promove a cooperação e a consequente valorização de todo e qualquer projecto.

Pretendemos ainda eu a criança se encontre envolvida num conjunto de relações, ou teia, que a levará de um conceito ao outro, proporcionando-lhe o acesso a toda a informação relevante de um modo fácil, transversal e adequado à apreensão do conhecimento e para isso teremos como interface do dicionário infantil uma plataforma *online*. Para além deste dinamismo que uma página *Web* pode proporcionar e de todas as vantagens didácticas que transporta, pode também engrandecer o espírito colaborativo da criança e tornar-se um instrumento importante para a construção de uma boa obra lexicográfica.

Faz parte ainda dos nossos objectivos organizar, rever e disponibilizar recursos para diversos fins. Através da recolha, da revisão e do tratamento de recursos *open source* estamos a contribuir para essa gigantesca tarefa de organizar os conteúdos da *Internet*, transformando-os em informação válida, contribuindo, desta forma, para assegurar, quer a qualidade do nosso trabalho, quer a possibilidade de todo este material ser usado para outros projectos que sigam a mesma política.

2. Metodologia

2.1. Prioridades

Para se dar início ao trabalho, como é apanágio em projectos desta natureza, sentimos a urgência de identificar quais as necessidades imediatas.

Era crucial fazer um pouco de "espionagem", ou seja, analisar alguns dicionários infantis existentes, tanto os de língua portuguesa, como os de línguas estrangeiras, nomeadamente os de língua francesa e inglesa, e tentar, a partir daí, identificar quer os pontos mais positivos, no sentido de alguma forma os adaptarmos ao nosso dicionário, quer as questões menos positivas a serem evitadas[1].

Identificámos como principal problema o carácter redutor das definições de grande parte deles, sendo que, por vezes, apenas são apresentados os sinónimos correspondentes a cada termo. Identificámos já aqui algumas daquelas que teriam de ser as nossas preocupações centrais: um cuidado especial na elaboração das definições, que teriam de ser simples e explicativas, visto que se destinavam a um público infantil e constituíam a maior valia para todo o nosso trabalho, e a aposta clara nos exemplos e abonações, pois um bom exemplo de uso, desde que bem pensado, transporta consigo informação importante e pode ser crucial para a explicação de um conceito.

Era necessário encontrar uma base consistente que funcionasse como ponto de partida e para isso era importante descobrir a maneira mais rápida e fácil de o fazer. Depois de alguma pesquisa na *Internet* conseguimos reunir alguns recursos relevantes, provenientes de glossários, de *Thesaurus* e de *stardicts*[2]. No entanto, decidimos adoptar como fonte principal uma base terminológica construída por um antigo estudante de LEA, Luís Gomes, que recorreu a um Stardict de Chinês/Inglês e decidiu fazer um dicionário de Chinês/Inglês/Português[3], à qual chamamos *TreeDic*. Escolhemos a *TreeDic* como base terminológica, porque possuía uma quantidade assinalável de termos aos quais estava já associado um conjunto relevante de relações semânticas. Esta decisão implicou que todos os outros recursos encontrados seriam considerados complementos da *TreeDic* e que seriam estudados e organizados em torno da mesma, no sentido de a completar.

Houve também a necessidade de vincular determinados objectivos, no sentido de organizar tarefas e metas a atingir para enquadrá-las dentro das nossas competências e do tempo disponível. Assim, ficou desde logo definido que todo o material recolhido nos diversos recursos seria organizado com a preocupação de ser projectado em vários formatos. Haveria então a preocupação central de conceber todo o nosso trabalho no sentido de poder ser projectado tanto numa plataforma *online,* tirando partido da flexibilidade e da explicabilidade do formato digital, como em *stardicts* e listas textuais, fornecendo assim um recurso renovado e organizado que poderá servir a outros fins. É possível, a partir desta estrutura, obter um resultado que possa ser impresso em formato papel, sempre mais inflexível, mas também ele importante. Ficou também prometido que, logo que possível, procederíamos à elaboração de uma página oficial que, mesmo

[1] Não fazemos referência a nenhum dicionário em particular uma vez que a breve análise feita teve como objectivo avaliar o material existente e identificar quais as necessidades maiores do público-alvo, não sendo objecto deste trabalho uma análise crítica dos dicionários infantis existentes.

[2] O *Stardict* é uma poderosa ferramenta com a qual podemos construir, descarregar e consultar dicionários. Está disponível para ser descarregada em http://stardict.sourceforge.net/ e mostrou-se ser uma ferramenta muito válida para quem precisa de trabalhar com dicionários.

[3] O método utilizado foi: traduzir para o Português o correspondente do Inglês obtido a partir do Chinês.

que de forma precária, fosse já apresentando os resultados obtidos e decidimos, mesmo antes de se apresentar a página propriamente dita, atribuir-lhe um nome: *Dicionauro*[4].

Este ponto da situação inicial foi absolutamente necessário porque identificou uma linha de orientação a seguir e permitiu que todo o trabalho fosse realizado no sentido de se atingirem determinados objectivos.

2.2 Recursos

Analisemos agora os recursos encontrados e examinados tendo em vista a reunião de uma lista considerada de termos para ser aplicada como macroestrutura[5] da *TreeDic*. Os recursos reunidos foram: a já referida *TreeDic*, que contava com cerca de 10.000 entradas; um *Freedict* de Inglês/Português (com cerca de 9.000 entradas); um dicionário dividido em duas partes: o *mikeharland* (+/- 3000 termos) e o *mikeharland2* (+/- 1200 termos); um glossário de termos sobre o tema *Internet* (180 entradas); um glossário de termos religiosos (58 entradas); uma lista de palavras frequentes do Português (300 entradas); um *Stardict* de Português/Inglês (5.334 entradas); um *Stardict* de Inglês/Francês (20.086 entradas); uma lista de sinónimos do Português (13.170 palavras – 4.002 grupos de sinónimos) e um *Thesaurus* (entre 4.600 e 4.900 entradas).

Como já foi referido, a *TreeDic* foi, desde o primeiro momento, a base de sustentação para compormos a nossa macroestrutura, quer por ser o recurso que já vinha da fase anterior, quer porque servia perfeitamente as nossas intenções. Assim, todos os outros recursos serão avaliados e revistos tendo em conta a estrutura e a arrumação da *TreeDic*, ou seja, tomaremos esta base terminológica como referência de revisão para todos os outros recursos, na esperança de virem a completá-la, pelo que aqui se fará a separação entre essa directoria base e os outros recursos.

2.2.1. Organização da base terminológica *TreeDic*

A *TreeDic* estava compilada numa pasta (*DICI*) dividida em vários ficheiros, que por sua vez estavam organizados por domínios, aos quais estava associado o vocabulário respeitante a esse domínio (área de conhecimento, actividade, etc.), vocabulário esse que estava apresentado em forma de ficha, possuindo uma série de informações e algumas relações. Vejamos o exemplo (1) para melhor visualizarmos o formato das fichas:

[4] Identificado o nome e por uma questão prática, sempre que nos referirmos ao nosso projecto iremos usar o termo Dicionauro.

[5] Entendemos a *macroestrutura* tal como a entende Rey-Debove, *apud* Iriarte Sanromán (2001, p. 24-25), ou seja, «l'ensemble des entrées ordonées, toujours soumise à une lecture verticale partielle lors du repérage de l'object du message».

(1)	PT espectáculo	(2)	PT nascente	(3)	PT actor
	-catgra		-catgra		-catgra
	-exuso		-exuso		-exuso
	EN show		EN source lake		EN actor
	CN 演出		CN 河源湖		CN 演□
	NU 600		NU 492		NU 3554
	Dom teatroBT		Dom rios		Dom cinema
	Def		BT		BT
			Def		Def

Nestes exemplos apresentados, verificamos que temos para cada entrada uma estrutura de relações orientadas ao conceito apresentado. O "**PT**" corresponde ao conceito (representado por um termo) português, ao qual são associados atributos: os atributos do termo, "**-catgra**" (categoria gramatical) e "**–exuso**" (exemplo de uso); os atributos do conceito: "**Dom**" (domínio), o "**BT**" (*broader term*) e o "**Def**" (definição do conceito). Ainda de referir os campos referentes às línguas: o "**EN**" (termo em Inglês), o "**CN**" (correspondente do Chinês). O "**NU**" representava o número que já estava atribuído por Luís Gomes e que se referia ao Chinês e que decidimos manter quer porque nos podia ajudar a identificar a ficha e o conceito a ela associada, quer porque, havendo necessidade, era sempre possível recuperar o chinês a partir desse "**NU**". Era esta então a estrutura que a *TreeDic* apresentava para cada ficha.

Para iniciarmos o nosso trabalho no que respeita a este recurso foi necessário traçar uma linha de acção, a qual seguisse a seguinte ordem:

(1) Proceder a uma revisão de todas as fichas existentes na *TreeDic*. Era fulcral saber em que estado se encontravam, visto a *TreeDic* ser um recurso que, para além de ter sido originalmente trabalhada para outros propósitos, tinha ainda sido posteriormente trabalhada por alunos da Universidade do Minho (nos quais nos incluímos), no âmbito da Disciplina de Bases de Dados Lexicais e Bibliotecas Digitais, ministrada pelo professor José João Almeida, com a colaboração do Professor Álvaro Iriarte Sanromán e do professor Alberto Simões. Esse trabalho, embora muito importante e relevante, apresentava alguns problemas, pelo que era agora necessário fazer uma triagem no sentido de analisar até que ponto esse trabalho executado poderia ser utilizado no nosso trabalho;

(2) Identificar, na totalidade das fichas, quais aquelas que seriam para manter e quais aquelas que não interessavam;

(3) Proceder desde já à revisão dos domínios[6], no sentido de identificar quais aqueles que já existiam e quais aqueles que pretendíamos modificar ou acrescentar.

[6] Identificámos que seria importante acrescentar a este trabalho informação de frequência em relação aos vários domínios incorporados e apresentar uma lista desses domínios. Entendemos que, devido ao carácter experimental e académico deste trabalho e à necessidade de hierarquizar prioridades, deveríamos nesta fase preocuparmo-nos com o lançamento das bases daquilo que poderá ser o *Dicionauro* e depois debruçar-nos--emos sobre a consolidação de todo o nosso projecto, da qual farão parte essas informações.

Para tal, foi necessário adoptar um método coerente que servisse para todos os recursos revistos e que permitisse agilizar também o tratamento informático dos mesmos, acelerando assim determinados processos de tratamento da informação.

Assim, procedeu-se a uma revisão destas 10.000 entradas, num processo demorado e complexo. No decorrer desta revisão, encontrámos uma série de problemas, sobre os quais era necessário reflectir e tomar uma decisão. Verificámos que grande parte das fichas tratadas no âmbito da disciplina de Bases de Dados Lexicais e Bibliotecas Digitais possuía problemas, nomeadamente, ao nível das definições (grande parte delas foram retiradas de dicionários gerais e não possuíam qualquer preocupação com o público-alvo), apresentando um grau de completude muito baixo, pelo que era necessário revê-las uma a uma, modificá-las e completá-las. Identificámos ainda que era necessário "arrumar" os sinónimos (atribuir-lhe uma arrumação). Havia ainda fichas que não possuíam domínio, sendo necessário identificá-las como pertencendo a uma qualquer área de conhecimento. Ao nível das correspondências entre o Inglês e o Português era também necessário revê-las uma a uma. Por fim, de salientar que um grande número de entradas levantou algumas dúvidas quanto ao grau de infantilidade e à sua pertinência para o *Dicionauro*.

2.2.2. Outros recursos

Foi adoptada uma estrutura de análise dos restantes recursos recolhidos, no sentido de uniformizar e de normalizar todas as revisões que fossem feitas e que ajudariam a completar a base terminológica *TreeDic*. No exemplo (2) é possível verificar a estrutura adoptada:

(4) COM:: PT:: EN:: DOM:: BT

O "**COM"** significa *comentário*, sendo que apenas se regista se for necessário e utilizando uma simbologia convencionada. O "**PT**" e o "**EN**" referem-se às línguas que estão presentes nesses recursos (que podiam ser outras). O "**DOM**" e o "**BT**" referem-se ao domínio e ao *broader term*, respectivamente, e são termos classificativos que explicaremos mais a frente, aquando da apresentação da nossa microestrutura. Veja-se de seguida o exemplo (3) para que se perceba melhor em que consiste esta normalização de revisões:

(5) =:: batata:: potato:: alimento:: tubérculo
(6) =:: bife:: steak:: alimento:: carne

O comentário (=) aparece em primeiro lugar, depois segue-se o termo em Português, o termo em Inglês, o domínio e o *broader term* correspondentes. Assim, identificámos logo quais as necessidades do recurso e quais as suas mais-valias para o *Dicionauro*, adoptando esta política para todos os recursos que foram objecto de revisão, exceptuando a *TreeDic* que nos serve de base.

No que se refere à lista de sinónimos do Português, ela foi retirada do *OpenThesaurusPT*, um projecto *open source*, com uma página *Web* interactiva, para a

criação de um dicionário de sinónimos na língua portuguesa. Com cerca de 13.170 termos (4.002 grupos de sinónimos), considerámo-lo ser um recurso já bastante organizado e bastante relevante para procedermos a alguma *arrumação*. Com a utilização deste recurso, para além de ganharmos mais algumas centenas de termos, também pudemos extrair informação relevante ao nível das relações semânticas de equivalência (sinonímia), pois a questão da sinonímia é uma questão bastante complexa no que respeita à concepção e organização de um dicionário.

Achamos também importante falar do *Thesaurus* da Unesco. Um *Thesaurus*, como refere Ana Lúcia Matos dos Reis, "é uma ferramenta da linguagem artificial de um domínio do conhecimento, construído por especialistas, especificando as relações entre os conceitos […] são úteis não só a profissionais da informação, mas também a cientistas, tradutores, engenheiros, especialistas [...]. O *Thesaurus* é utilizado na indexação das informações (entrada de dados numa base) e na recuperação da informação (saída de dados de uma base), independentemente da área do saber em que se enquadre" (2001: 4). Este *Thesaurus* possui entre 4600 a 4900 termos, os correspondentes de grande parte desses termos em quatro línguas: Espanhol, Inglês, Francês e Português, e uma rede de relações, tais como: NT (*narrow term*), BT (*broader term*), RT (*related term*) UF (*used for*) SN (*definition*) MT (*term*). Estas relações, à semelhança de grande parte dos *Thesauri* actuais, são de três tipos: relações de equivalência ou substituição (sinonímia), de hierarquia (hiperonímia e hiponímia) e de associação (meronímia).

Notámos, numa primeira análise, que alguns dos termos presentes neste *Thesaurus* eram referentes a áreas como a economia, assuntos estrangeiros, etc., o que limita o seu uso para efeitos de elaboração de um dicionário infantil. No entanto, o conjunto de relações que possui é de extrema relevância e pode ser utilizado como modelo para a construção do *Dicionauro*.

De salientar ainda que todos os recursos utilizados teriam uma codificação estandardizada, convencionando-se que todos seriam codificados como UTF-8, "um tipo de codificação *Unicode*[7] que está a ser adoptado como tipo de codificação padrão para *emails*, páginas *Web*, e outros locais onde os caracteres são armazenados, e que pode representar qualquer carácter universal do padrão do *Unicode*" (http://pt. wikipedia.org/wiki/Utf-8). Isto para que todos os caracteres usados nos nossos documentos pudessem ser vistos em qualquer computador.

Todos os recursos aqui referenciados e que se mantiveram como objecto de análise, foram e continuam a ser objecto de constantes revisões no sentido de os tornar mais ricos e mais organizados, quer para serem utilizados no nosso *Dicionauro*, quer para serem disponibilizados em listas textuais e assim servirem a projectos que a eles queiram recorrer, dando desta forma mais força à política de desenvolvimento por nós adoptada: identificar as carências dos recursos recolhidos, organizá-los mediante as nossas necessidades e disponibilizá-los para que, também eles possam ser um recurso *open source* melhorado.

[7] Para a noção de *Unicode vd*. http://www.unicode.org/.

3. Microestrutura

Superada esta fase de colecção de material, de descrição das prioridades e de revisão dos recursos que servirá de base à nossa macroestrutura, deparámo-nos com a necessidade de organizá-la e ajustá-la às necessidades deste projecto. Era então necessário encontrar uma estrutura sólida e adaptada, quer às necessidades de apreensão e de compreensão do conhecimento do público infantil, quer às necessidades informáticas de automatização do processamento dessa informação, quer ainda ao compromisso de partilha e disponibilização melhorada dos recursos coleccionados.

Para responder a essas necessárias e tendo como base algumas das relações que já figuravam na *TreeDic*, assim como alguma da política de estruturação herdada da arrumação feita na sequência da disciplina de Base de Dados e Bibliotecas Digitais, vejamos o exemplo (4), que ilustra a microestrutura[8] adoptada:

(7) **PT** cão
PT_BR cachorro
-fem cadela
-catgra nm
-exuso o cão do João ladrou quando ouviu um barulho
AUDPT cao.mp3
EN *dog*
AUDEN *dog.mp3*
FR *chien*
AUDFR *chien.mp3*
ES *perro*
AUDES *perro.mp3*
DE *hund*
AUDDE *hund.mp3*
CN
NU *562*
Dom *animal*
BT *animal doméstico*
NT *canídeo*
RT *canino*
PART *focinho*
POF *matilha*
Def *animal doméstico que ladra (...)*
IMG *cão.jpg*
UP *usado para guardar a casa, caçar (...)*
GI *4*
IM *4*
ID *manuel*

[8] Entendemos a microestrutura tal como a entende Rey-Debove, *apud* Iriarte Sanromán (2001, pp. 24-25), ou seja, «l'ensemble des informations ordenées de chaque article, réalisant un programme d'information constant pour tous les articles, et Qui se lisent horizontalement à la suite de l'entrée».

O primeiro campo representa o conceito em Português (***PT***), representado por um termo que, no caso do exemplo apresentado, é o termo "*cão"*. Este PT refere-se aos termos que são comuns ao Português europeu e ao Português do Brasil, o PT_PT aos termos apenas do Português europeu e o PT_BR apenas à variante do Português do Brasil. De referir ainda que decidimos que o nosso dicionário fosse orientado ao conceito, ou seja, cada termo corresponde, ou representa um conceito. Tanto a unidade palavra, como as palavras compostas, como as palavras derivadas, como as unidades *pluriverbais*, os *pragmatemas* e as denominações perifrásticas, etc., usadas aqui como unidades lexicais, serão tidas como conceitos, não obstante o facto de quando se fizer a projecção em formato papel se tenha a preocupação de arrumar devidamente todo este material.

Os atributos do termo pretendem fornecer informação gramatical e exemplos concretos do uso das palavras: a categoria gramatical (*-catgra*) a que o termo pertence, e, sempre que pertinente (no caso das formas irregulares), alguma informação morfológica: género (*-fem/masc – cão/cadela*), ou número (*-plural – cão/cães*), e exemplos de uso (*-exuso*). Esta última informação assume uma importância capital em todo o nosso trabalho. Os exemplos de uso são, na nossa opinião, uma mais-valia quando falamos de dicionários gerais e mais ainda quando tratamos de dicionários infantis, ou dicionários de aprendizagem, porque "podem ser muito ricos em informação gramatical, enciclopédica, pragmática ou sobre combinatória lexical" (Iriarte Sanromán, 2001: 327) e porque achamos que podem ser aproveitados como complementos às definições dos conceitos. Para isso foi necessário evitar que esta informação se transformasse perigosamente numa espécie de "caixote do lixo para onde mandamos tudo aquilo que não sabemos tratar lexicograficamente" (*idem*, 328) e, também, acautelarmo-nos para o perigo de tais exemplos apenas servirem para aumentar o número de informação que é fornecida no dicionário e não acrescentarem informação relevante. Procurou-se sempre adequar o exemplo de uso às necessidades explicativas do conceito de forma a introduzir informação importante e esclarecedora, podendo mesmo funcionar como uma forma de o definir.

Para os exemplos de uso recorremos na maioria das vezes à intuição linguística e à subjectividade, mas também aos dicionários já existentes, a alguns manuais escolares e ainda a algum *corpus* disponível na *Internet*[9]. Como refere Iriarte Sanromán, "os lexicógrafos não podem continuar a trabalhar sem tomar como ponto de referência um *corpus* […] que lhes permita fazer generalizações em relação a usos" (2001: 112). Concordamos plenamente com esta afirmação pois será esta prática que atribuirá ao nosso trabalho a automatização e a base real que nos permita justificar as nossas escolhas. No entanto, se pretendemos adequar os exemplos de uso às necessidades explicativas que cada conceito apresenta, temos de analisar caso a caso e recorrer na maioria das vezes à á referida intuição linguística e à subjectividade, num processo demorado e complexo.

[9] *Corpus* CETEMPúblico (http://www.linguateca.pt/CETEMPublico)

Em relação aos atributos do conceito, temos uma série de informações que são um misto de relações e associações importantes, que o envolvem numa teia explicativa que facilita a sua compreensão. Decidimos incluir na nossa microestrutura os seguintes atributos referentes ao conceito: um domínio (*Dom*), um *broader term* (*BT*), um *narrow term* (*NT*), um *related term* (*RT*), um *part* (*PART*), um *part of* (*POF*), um *used for* (*UP*) e uma definição (*Def*).

Este conjunto de campos segue uma tendência moderna de estruturação no que diz respeito ao processamento de linguagem natural, sem ofender os princípios básicos de elaboração de dicionários. Esta microestrutura segue esse objectivo claro da necessidade de construir um vocabulário que possa ser entendido por uma comunidade, que no nosso caso particular é o público infantil e que possa também ser compreendido, compartilhado e manipulado pelos agentes de inteligência artificial[10]. Com estas relações pretende-se que os termos sejam usados para descrever as várias áreas do conhecimento e que, assim, se construa a sua representação. Isto caracteriza aquilo que hoje se designa nas ciências da computação por "ontologia": um "conjunto de termos ordenados hierarquicamente para descrever um domínio que pode ser usado como um esqueleto para uma base de conhecimento", (http://www2.dbd.puc-rio.br/pergamum/tesesabertas/0024134_02_cap_04.pdf). A esta noção de ontologia está associada a de *Thesaurus*[11].

Existe uma dificuldade muito grande em estabelecer fronteiras entre aquilo que são as ontologias e o *Thesaurus*, considerando-se muitas vezes que os *Thesauri* são "ontologias mais simples". Se é verdade que numa acepção moderna a ontologia é tida como essa "disciplina que estuda e determina as relações entre os conceitos estabelecendo regras lógicas de raciocínio sobre esses conceitos gerando linguagens que são compreendidas pelos computadores" (http://www.sibi.ufrj.br/snbu/snbu2002/oralpdf/124.a.pdf), também é verdade que à ideia de *Thesaurus* está associada uma ideia de sistema, um sistema composto por conceitos e que são representados por termos, sendo que cada termo tem obrigatoriamente uma ligação com outro termo, ou outros termos, e é esta relação que estrutura o *Thesaurus* (Matos dos Reis, 2006: 6) e que "permite que se aceda à informação sob diferentes ângulos" (*idem*, 2).

É importante referir ainda que as relações de hierarquia existentes (que estão necessariamente contempladas na construção de *Thesaurus* e de ontologias), seguem uma taxonomia, de um nível topo, mais abrangente, para um nível mais específico. No nível mais elevado encontra-se o *domínio* (*DOM)*, a área de conhecimento a que o termo pertence, que carrega consigo informação de tipo enciclopédico-cognitiva; seguida do *broader term* (BT), que representa o nível que vem imediatamente a seguir ao *domínio* na hierarquia de relações do Dicionauro (se o *Dom* é uma área de conhecimento geral, o *BT* irá apontar uma área mais específica dentro dessa área de

[10] *Vd.* SOUZA, R.R., ALVARENGA, L., (2004). A Web semântica e suas contribuições para a ciência da informação. Acedido em 9 de Agosto de 2008 no sítio da Internet da Scientific Electronic Library Online: http://www.scielo.br/scielo.php?pid=S0100-19652004000100016&script=sci_arttext&tlng=pt

[11] Para mais informação acerca dos *Thesaura vd.* www.elprofessionaldelainformacion.com/contenidos/1994/febrero/consideraciones_sobre_los_Thesaura.html

conhecimento); do *narrow term* (*NT*), a categoria mais baixa e que se encontra mesmo abaixo do próprio termo; do *part* (PART) e o *part of* (POF). Estes dois últimos são os campos responsáveis por estabelecer essa relação de hierarquia semântica entre os conceitos: um denota a parte (PART) e o outro denota o todo (POF) – relações de meronímia e holonímia. Vejamos um exemplo:

(8) PT fruto
PART semente
PART polpa
POF árvore

Neste caso a unidade lexical "fruto" (a parte – merónimo) implica a unidade lexical "árvore" (o todo – holónimo), isto para o caso do POF. Para o campo PART a unidade lexical "polpa" (a parte – merónimo) implica a unidade lexical "fruto" (o todo – holónimo). Assinala-se aqui o carácter simétrico destas duas relações, sendo que a sua leitura permite que se aceda à informação sobre diferentes ângulos, atestando aqui a importância dos tesauros para a organização dos documentos e do conhecimento.

Para além destas categorias, que estabelecem relações hierárquicas na nossa árvore, existem também as relações de associação, um conjunto de relações livres no âmbito do *Thesaurus*, que na nossa microestrutura aparecem identificadas como *related term* (*RT*) e que inclui todos os termos relacionados com o conceito que se está a descrever (antónimos, homónimos, parónimos, etc.

Detenhamo-nos agora sobre a definição (Def). A definição é a parte mais importante de um dicionário orientado a um público infantil e, no que respeita aos dicionários infantis que consultámos, é aquela a que se dá menos importância aquando da sua elaboração. Durante este tempo e quando definíamos os conceitos tentámos combater essa tendência de "redução" das definições dos dicionários gerais a que maior parte das vezes se recorre para elaborar os dicionários infantis e construir definições simples e apropriadas ao público infantil. Existem vários métodos para encontrar a definição que melhor se adapta à criança: a paráfrase, um sinónimo, um exemplo pragmático, como vimos atrás, a associação à classe (uma explicação *naïve* – a libandisca é um pássaro), ou ainda recorrendo à antonímia. Coube-nos a nós encontrar a melhor forma de definir o conceito e foi nesse processo de análise e de introspecção que encontrámos o melhor mecanismo, aquele que se adapta perfeitamente, mediante as necessidades de explicação do conceito. É sempre necessário ter algum cuidado na construção das definições para que elas constituam essa forma simples e intuitiva de apreensão e compreensão dos conceitos. Foi com este cuidado que procuramos construir as nossas definições.

Associada à definição podemos referir o campo UP (usado para). O UP corresponde a mais alguma informação relevante, do tipo pragmático-contextual e retórica, que já poderá estar contida no termo ou não, e que alarga tanto a informação relativa à definição do conceito, como as funcionalidades do nosso recurso.

Por fim, deparemo-nos sobre os restantes campos contemplados na nossa microestrutura que ainda não foram abordados e que convém aqui referir. São eles IMG

(imagem), AUD (áudio), GI (grau de infantilidade) e IM (índice de maturidade). Os dois primeiros correspondem a campos de ilustração que pretendemos que estejam contemplados no nosso dicionário: a imagem e o formato áudio. Sempre que for pertinente associaremos ao conceito uma imagem ilustrativa, pois concordámos que as imagens são um complemento importante para este tipo de dicionário, e a cada campo referente às línguas, incluindo o Português, um ficheiro áudio (AUDPT, AUDEN, AUDFR, AUDES e AUDDE), que corresponde à gravação áudio do termo em questão, que apenas será pertinente para o formato electrónico do nosso dicionário e que pretende fornecer informação fonética acerca das línguas em questão ou alargar o dicionário a um público mais abrangente (os cegos, por exemplo). O GI e o IM são campos que decidimos incluir para termos uma noção do estado de completude da ficha e da sua adaptabilidade para o público-alvo.

4. Trabalhos futuros e considerações finais

A disponibilidade e a crescente facilidade de acesso aos computadores e à *Internet* tem vindo a crescer exponencialmente nos últimos anos e tem contribuído para trazer profundas implicações no que diz respeito à preparação de um dicionário e até à distribuição dos seus conteúdos. Desde alguns anos a esta parte, os lexicógrafos, com a ajuda de programadores informáticos, têm vindo a criar dicionários, *Thesaurus* e outros trabalhos de referência, partindo de bases de dados codificadas. É importante, ainda, que os profissionais das letras possam adquirir conhecimentos na área da informática, não só em linguagens de anotação, mas também em disciplinas específicas que os ensinem a programar de acordo com as suas necessidades, porque a lexicografia moderna obriga-nos a desafiar estas dificuldades e a admitir a importância da lexicografia computacional.

Este trabalho foi e continuará a ser um desafio para nós. Desde o processo de levantamento e de revisão dos recursos até às projecções feitas (listas textuais, plataforma *online*), tudo teve de ser pensado para cobrir dois objectivos: o de adequar a um público infantil todo o material recolhido e o de o organizar tendo em vista o processamento informático. Este método resulta de um esforço da nossa parte em aproveitar essa grande capacidade de armazenamento, de recuperação e de tratamento exaustivo de grandes quantidades de informação que a informática nos permite actualmente e aliá-la aos nossos conhecimentos ao nível da linguística, particularmente da lexicografia. Apesar de muitas vezes frustrado, visto que tivemos de mudar muitas vezes os métodos e a estrutura da nossa base, consideramos que foi positivo, na medida em que, percorrendo um caminho cheio de obstáculos, aprendemos a estar atentos e a ter uma visão crítica em relação a toda a informação que utilizamos, desde os recursos às traduções.

Por fim, queremos dizer que apesar de toda a satisfação que sentimos pelo trabalho realizado e por termos conseguido atingir os objectivos a que nos propusemos nesta primeira fase, ainda há muitas arestas a limar, ainda há muito caminho a percorrer neste trabalho de revisão, organização e partilha, pois nunca se pode considerar um instrumento da língua algo completo. A língua vai sofrendo alterações ao longo do tempo e o dicionário deve ser o registo dessa mudança.

5. Referências

Haensch, G., L. Wolf, S. Ettinger & Werner (1982) *La Lexicografía. De la lingüística teórica a la lexicografía práctica.* Madrid: Gredos.

Iriarte Sanromán, Álvaro (2001) *A unidade lexicográfica. Palavras, Colocações, Frasemas, pragmatemas.* Braga: Centro de Estudos Humanísticos – Universidade do Minho.

Iriarte Sanromán, Álvaro (2004) Dicionários codificadores. In C. M. de Sousa e R. Patrício (2004) *Largo Mundo Alumiado. Estudos em Homenagem a Vítor Aguiar e Silva.* Braga: Centro de Estudos Humanísticos – Universidade do Minho.

Moreira, Alexandra (2003) *Tesauros e ontologias: estudo de definições presentes na literatura das áreas das ciências da computação e da informação, utilizando-se o método analítico-sintéctico*. Programa de Pós-Graduação em Ciências da Informação. Belo Horizonte: Escola de Ciências da informação da Universidade de Minas Gerais. 151 pp. [em linha]. [consult. em 15 de Agosto de 2008]. Disponível em: http://opus.grude.ufmg.br/opus/opusanexos.nsf/4d078acf4b397b3f83256e86004d9d55/915f0db8ceb5bb3583256fb0006a1d5e/$FILE/mestrado%20-%20Alexandra%20Moreira.pdf

Matos Dos Reis, Ana Lúcia (2006) *Os tesauros e as vantagens do formato XML*. Pós Graduação em Ciências da Informação e da Documentação, Tecnologias de Informação documental. Porto: Universidade Fernando Pessoa. 14 pp. [em linha]. [consult. em 10 de Agosto de 2008]. Disponível em: (http://www.cerem.ufp.pt/~nribeiro/aulas/tid/TID_Ana_Lucia_Reis.pdf.)

Matos Dos Reis, Ana Lúcia (2006) *4.Ontologia.* Acedido em 20 de Agosto, de 2008, em http://www2.dbd.puc-rio.br/pergamum/tesesabertas/0024134_02_cap_04.pdf

Oliveira, Rosa Maria Vivona Bertolini (sd) *Web Semântica – novo desafio para os profissionais da informação* [em linha]. [consult. em 20 de Agosto de 2008]. Disponível em: http://www.sibi.ufrj.br/snbu/snbu2002/oralpdf/124.a.pdf.

Souza, R.R. & L. Alvarenga (2004) *A Web semântica e suas contribuições para a ciência da informação* [em linha]. [consult. Em 9 de Agosto de 2008 no sítio da *Internet* da Scientific Electronic Library Online]. Disponível em: (http://www.scielo.br/scielo.php?pid=S0100-19652004000100016&script=sci_arttext&tlng=pt)