"Less, Easier and Quicker" in Language Acquisition for Patent MT

Svetlana SHEREMETYEVA
LanA Consulting
Mynstersvej, 7a
Copenhagen, Denmark, DK-1827
lanaconsult@mail.dk

Abstract

The paper describes some ways to save on knowledge acquisition when developing MT systems for patents by reducing the size of resources to be acquired, and creating intelligent software for knowledge handling and access speed. The approach is illustrated by knowledge acquisition and maintenance in the APTrans system for translating patent claims. Domain tuned resources are based on contrastive studies of multilingual patent documents and are handled by an electronic dictionary with a powerful user-friendly environment for acquisition, editing, browsing, defaulting and coherence proofing.

1 Introduction

Acquiring and handling language resources for high quality machine translation is very time consuming. It is therefore imperative to find ways of facilitating and speeding up acquisition/handling of useful static knowledge sources.

There are at least two directions of work that can contribute to the problem.

The first one is to reduce the amount of language acquisition to a "must" minimum for a given application, - it is well known that the complexity of developing full-sized knowledge bases for various NLP systems makes reusability of such resources highly desirable, but this process is extremely expensive and time consuming, and reusability is not guaranteed and knowledge contributing to adequate performance even of a single NLP system is already useful (Cowie and Lehnert 1996).

The question of what linguistic information should be stored in a lexicon is very important and does not have a single answer. Massive attempts have been made to work out general statements on the content of the lexicon, ranging from prosodical patterns of the word to its detailed semantic characteristics and usage (Chomsky 1995; Atkins and Levin 1991; Hudson 1995). The Computational Lexicon Working Group of the Text Encoding Initiative[1] suggests standards for interchanging lexicon data intended for use by NLP systems. While this trend is clearly important, practically any viable application developed within a reasonable period of time has a lexicon with the content and encoding constrained by the purpose for which it is built. The main thing is to specify what is a must for a particular application (patent claim MT, in our case) tuned lexicon.

The second way to speed up acquisition is to develop intelligent knowledge handling tools, of which electronic dictionaries are most essential.

In this paper we try to contribute to the problem with a case study, - TransDict, - a multilingual domain tuned lexicon being developed for a family of patent-related NLP applications, such as AutoPat, APTrans and AutoRead[2]. TransDict thus conforms to the "Multilingual-Specialized" dictionary paradigm (Sérasset, 1993) and features a powerful environment for acquisition, editing, browsing, defaulting and coherence proofing.

TransDict is implemented in C++ as an integral part of 32-bit Windows applications for the: Windows 95/98/2000/NT operating environments. The languages that are currently being covered are English, Danish and Russian but TransDict can easily be extended to a multiple number of other languages and domains.

In what follows we first describe the methodology of the TransDict content specification and then describe the TransDict tool features.

2 What is in a claim? Corpus analysis

The claim is the focal point of a patent disclosure and is the actual subject of legal protection. The claim describes essential features of the invention in the obligatory form of a single extended nominal sentence with a well-specified conceptual, syntactic and stylistic/rhetorical structure which frequently includes long and

[1] http://nl.ijs.si/et/Talks/asslli02/

[2] AutoPat, APTrans, AutoRead, - computer systems for authoring, translation and improving readability of patent claims, correspondingly (Sheremetyeva, 2003)

telescopic embedded predicate phrases. Figure 1 illustrates a fragment of a US claim text (claims can be over a page long).

A brush ***comprising*** *a handle* ***at*** *one end, a brush head* ***at*** *the opposite end, a plurality of bristle holders and a plurality of crank arms, each of said crank arms* ***being coupled*** *to one of said bristle holders, each of said bristle holders* ***is mounted*** *rotatably in said brush head, each of said crank arms* ***is formed*** *with an elongated slot and a member* ***mounted*** *in said handle, said member* ***carrying*** *a plurality of pins, each of said pins* ***is received*** *in said slot to oscillate a respective bristle holder.*

Figure 1. An example of a patent claim text. Predicative words which are heads of individual phrases describing essential features of the invention are bold faced.

APTrans confines to a general MT paradigm in that it includes analysis, transfer and generation and relies on corpus based lexical, morphological, syntactic and semantic knowledge. We illustrate our lexicon specification methodology on the example constructing a lexicon based a 9-mio-word corpus of US claims. We used a mixture of automatic and manual procedures with the elements of statistical analysis.

Corpus analysis included:

- Automatic acquisition of the corpus frequency list.
- Automatic suffix-based sorting of the initial word form list into "dirty" lists of parts-of-speech.
- Human aided sorting the "dirty" wordform lists into "clean" lists of parts of speech
- Human aided sorting words within every POS into semantic classes, relevant for particular parts of speech. A set of semantic classes thus extracted is in fact a coarse patent domain ontology.
- Human aided sorting the words within every POS into different morphological forms.
- Statistical and qualitative analysis of all the lists

The results showed that sublanguage of patent claims is very restricted, - a 9-mio-word corpus amounted to approximately 60.000 different word forms. Some of the words did not emerge in full paradigms, e.g., many nouns function in singular- or in plural form only; verbs can miss a lot of their forms.

To be robust the system lexicon should of course contain full paradigms of nouns and other words, as for verbs, their description can be more domain specific.

Verbs are predicative words whose properties are mainly responsible for the claim structure. Therefore we paid special attention to verbs-predicates (words that are bold faced in Figure 1). The results of their analysis are given below.

2.1 Predicates

Morphology. The results of statistical analysis of the corpus showed that the inventories of predicates in patent claims are very restricted. About 600 of predicates cover 98% of all predicate words forms in the corpus.

The grammatical forms in which these predicates occur in the corpus are also quite restricted: 92% of the text realizations of the predicates are covered by 7 most frequent forms. These are listed below in the descending order of frequency of occurrence in the corpus: 1. Past Participle (*mounted*); 2. Present Participle (*comprising*); 3. Present Simple Tense Passive (*is mounted*); 4. Present Simple Tense Active (*connect*s); 5. Infinitive Simple Active (*to move*); 6. Gerund Simple Active (*engaging*); 7. Present Participle Continuous Passive (*being held*).

The predicates display a rather strong lexical, morphological and syntactic correlation (see Table 1), e.g. the verb "include" is mostly used in the form of Present Participle, the verb "mount" is found most frequently in the form of Past Participle, etc.

To facilitate MT, passive and active surface forms of one and the same verb are treated as different words[3] and described by different dictionary entries (the forms of Present Participle and Past Participle are taken as canonical forms for active and passive predicates, respectively), for example, *formed* and *forming* are treated as different predicate lexemes.

For the same reason, predicates which are used in different senses (there still exist polysemantic predicates in the patent sublanguage, though to a much smaller extent compared with the general language) are decomposed into a corresponding number of homonymous single-sense predicates. For example, in the phrases *the element* ***mounted*** *on the base* and *the element* ***mounted*** *to the base* the predicate *mounted* means *positioned* and *connected* respectively.

[3] Actually, this technique has been used by a number of researchers and is supported by the psycholinguistic data showing that generation of passive structures is effected by principally the same mechanisms as active structures rather than by passivization transformations of active structures into passive (Byrne 1966; Johnson 1967).

N	predicate	Total Freq.	Past Participle	Present Participle	Present Passive Simple	Present Active Simple	Gerund Active Simple	Infinit. Active Simple	Infinit. Passive Simple	Other
1	have	565		433		122	5	5		
2	provide	360	27	141	108	20	12	39		13
3	comprise	254	6	184	4	60				
4	include	228		142	4	80				2
5	extend	206	3	140	2	48	1	7		5
6	mount	205	127	5	41	1	12	3	2	14
7	form	201	66	29	31	14	11	41	2	7
8	be	189		49		135		5		
9	support	148	28	49	23	5	25	13	3	2
10	connect	145	69	17	30	1	10		8	10

Table1: Correlation of lexical and grammatical forms of US predicates (a fragment)

FORM	Realization of Active predicates	Realization of Passive predicates	Realization of Adjectives
`full`	Present Participle Gerund Simple Active Infinitive Simple Active	Past participle	Adjective
`short`	Present Simple Active	Present Simple Passive	is/are + Adjective
`absolute`	Present Participle	Present Participle Continuous Passive	being + Adjective

Table 2: FORMs of US active and passive predicates

It is interesting that in the patent claims different voice forms of the same predicates sometimes Treating passive and active realizations of verbs as different predicates allows for a further reduction in the variety of grammatical forms of the predicates. Due to the restrictions on the sublanguage, surface realizations of the predicates can be represented as a fixed set of traditional grammatical categories, a FORM, having only three values, which we called full, short and absolute,for the US predicates and two values, full and short for Russian as explained in Table 2.

On analyzing the frequency vocabulary of the lexical, morphological and syntactic correlations among the predicates in terms of FORMs further sublanguage restrictions could be clearly seen: most of the verbs and adjectives preferably function in one of the FORMs (usually, `full` or `short`); moreover, if a major morphological form of an English active predicate is `full`, it is realized as Present Participle. For example, the predicate *forming* functions most frequently in the `full` FORM *forming,* the major FORM of the predicate *superimposed* is `short`*: is/are superimposed.* These results are used to develop heuristics simplifying analysis and generation procedure.

Semantic classes. In the patent sublanguage both legal and technical domain components restrict the selection and usage of language units. The predicates in patents can be classified into the following types: 1) meronymy; 2) property; 3) spatial; 4) connection; 4) interaction; 6) comparison; 7) change-location; 8) limitation; 9) separation; 10) purpose; 11) structural peculiarities, etc. To simplify MT, these classes are ranked in the order that they must appear in the patent claim text. An invention should be described by specifying, in this order: 1) its components (and components of components, as required); 2) properties ("attributes") of the components (shape, material, etc.); and 3) relations among the components (spatial, connection, etc.)

Case roles. To determine predicate-argument structures in the sublanguage, a valence analysis (Fillmore 1970) of the predicates on both semantic and syntactic levels was carried out by analyzing predicate distribution in the claim text corpora. The list of case roles for the sublanguage is defined as follows: *subject, object1, object2, place, manner, purpose, means,source, destination, parameter, condition, time.* The set of case-roles is not necessarily the same for every predicate and not all case-roles defined for a predicate co-occur every time it appears in the claim text. At this acquisition stage we also collected all sublanguage case-role fillers. This information will be used to simplify the morphological analysis of case role values supplied by the user. For example, the case role *purpose* in the sublanguage of the US claims can be realized as follows:

purpose: ((*for* Ger)(*for* N)(*for* NP)(*for the purpose of* Ger) Inf (*so that* S)(*so as* Inf)),

where N denotes a noun, NP, a noun phrase, Ger, a gerund, Inf, an infinitive, and S, a clause.

Linearization patterns. Linearization patterns of predicates contain knowledge about co-occurrences of predicates with particular case roles. The extraction of lexical co-occurrence knowledge has been the subject of a number of studies, e.g. (Calzolari and Bindi 1990; Church et al. 1991) and ranged from simple extraction of word associations from corpora to the extraction of word associations augmented with part-of-speech and semantic tagging.

The general motivation for the interest in co-occurrence knowledge is that it "can be helpful for lexical disambiguation in analysis and crucial for lexical selection in generation" (Calzolari and Bindi 1990: 58). The main peculiarity of co-occurrence knowledge in our system is that it was augmented with case-role tags, which include syntactic and semantic information.

Thus, for example, the following phrase from an actual claim: (1: *the splice holder*) *: *is mounted* (2: *on the cover part*) (4: *to form a rotatable splice holder*) (where 1, 2 and 4 are case role ranks and "*" shows the position of the predicate) will match the linearization pattern (1 * 2 4). We acquired a list of co-occurrence linearization patterns for the predicates' case frames and ranked them in the order of decreasing frequency.

We have described the corpus relevant linguistic knowledge, which in our MT system is stored, updated and handled by TransDict.

3 TransDict

Filling text, A vast amount of research in the field of electronic dictionaries concentrate on data unification, representation, organization and management with the major focus on multilingual dictionaries as, for example, in (Wong, 2000; Boitet et al.,2002). Multilingual electronic dictionaries most often include a database of cross-referenced unilingual dictionaries with the use of interlingua such as ontology (Onyshkevich and Nirenburg, 1994)) or a pivotal language (Boitet et al.,cf.).

The architecture of such dictionaries normally include a lexical database and a set of tools for data management, - visualisers, editors, defaulters, etc. (Khatchadourian, 1992), a user-friendly interface being one of the most important (Bilac and Zock, 2003). XML, and SGML data representation languages (Boitet et al., cf.) have been a successful approach to facilitate the export of electronic dictionaries to different applications though many dictionaries use their own internal data representation formats (Fedder, 1992).

Finally, it is desirable for electronic dictionaries to be stand-alone modules with defined interfaces for interaction with other linguistic applications (Pointer project report, http://www.computing.surrey.ac.uk/ai/pointer).

3.1 TransDict feature space

TransDict is built over the set of features relevant for the applications as cited above**:**

Semantic features: SEM_Cl - semantic class, CASE_ROLEs, - a set of case roles associated with a lexeme, if any).

Syntactic features: FILLERs, - sets of most probable fillers of case-roles in terms of types of phrases and lexical preferences.

Linking features: PATTERNs, - linearization patterns of lexemes that code both the knowledge about co-occurrences of lexemes with their case-roles and the knowledge about their linear order.

Morphological features: POS, - part of speech, MORPH, - wordforms, number, gender, etc.; the sets of parts of speech and wordforms are domain and application specific (Sheremetyeva, cf.).

Rank feature: RANK, - corpus-based frequency within one semantic class. The more frequent is a lexeme, the less its rank.

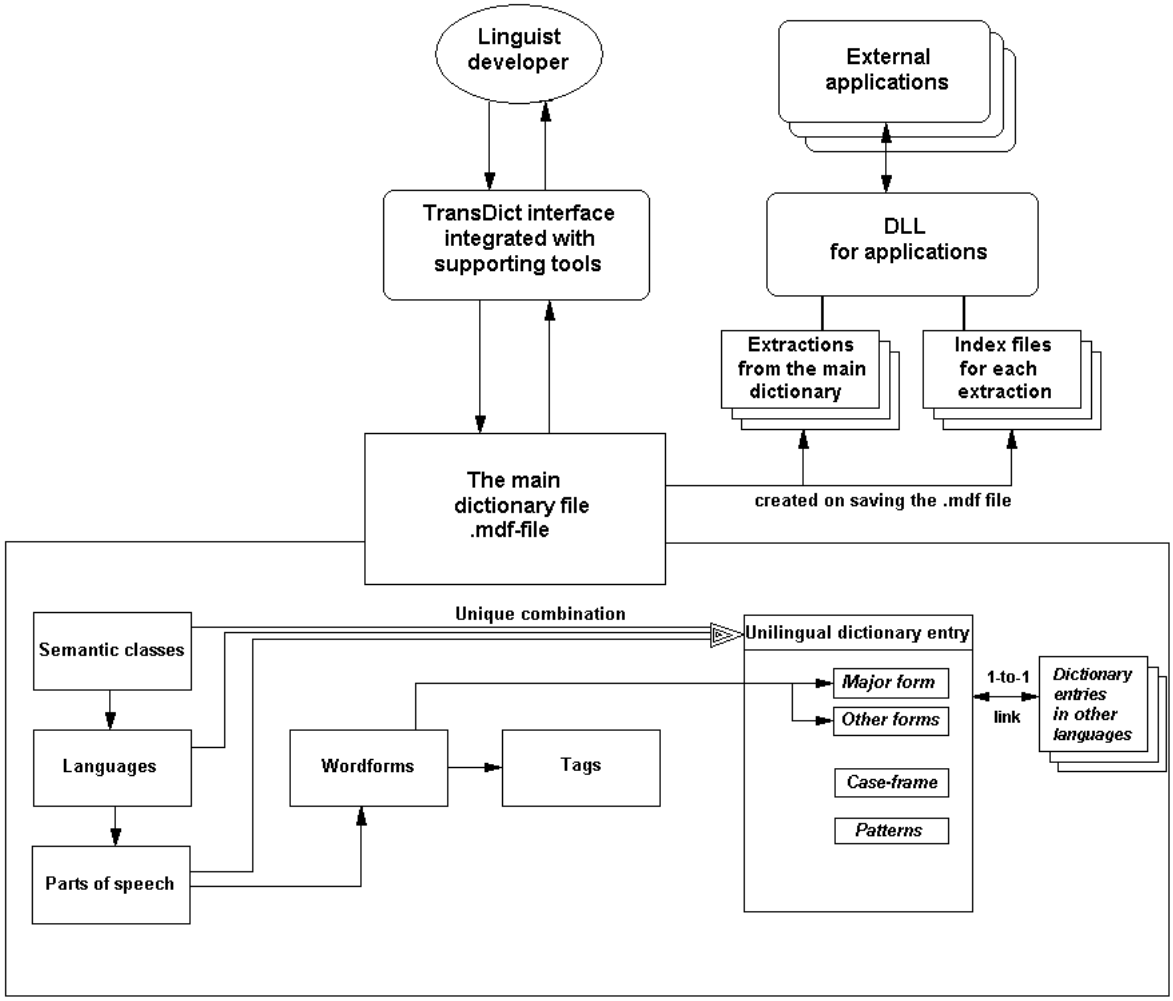

Figure 2. An overall architecture of TransDict.

3.2 Organization and architecture

TransDict includes cross-referenced monolingual lexicons for every language. A monolingual dictionary consists of a set of entries. An entry identifies lexical information for one meaning of a lexeme of a given language. Every entry is maximally defined as a tree of features:

SEM-CL[Language[POS RANK [MORPH CASE_ROLE FILLER PATTERN]
The CASE_ROLE , FILLER and PATTERN features might not be specified in certain entries, e.g., for nouns-physical objects.
A maximal entry has the following fields:

entry::=
semantics SEM_CL
language LANGUAGE
part of speech POS
major-form string TAG
other-forms {string TAG}+
case-frame {CASE_ROLE}+
filler {CASE_ROLE{FILLER}+}+
patterns {PATTERN}+
frequency RANK
translation{cross-linguistic equivalent entry index}+

TAG is a label, which codes several features, such as POS, number, inflection type and semantic class (physical object, substance, event, etc.).

The architecture of TransDict is shown in Figure 2. The developer works with the main dictionary file (MDF) visualised by the interface (Figure 3). All information is stored in TransDict internal formats: in data files and index files. When the lexicographer saves the data multiple extractions from the main dictionary file are automatically created.

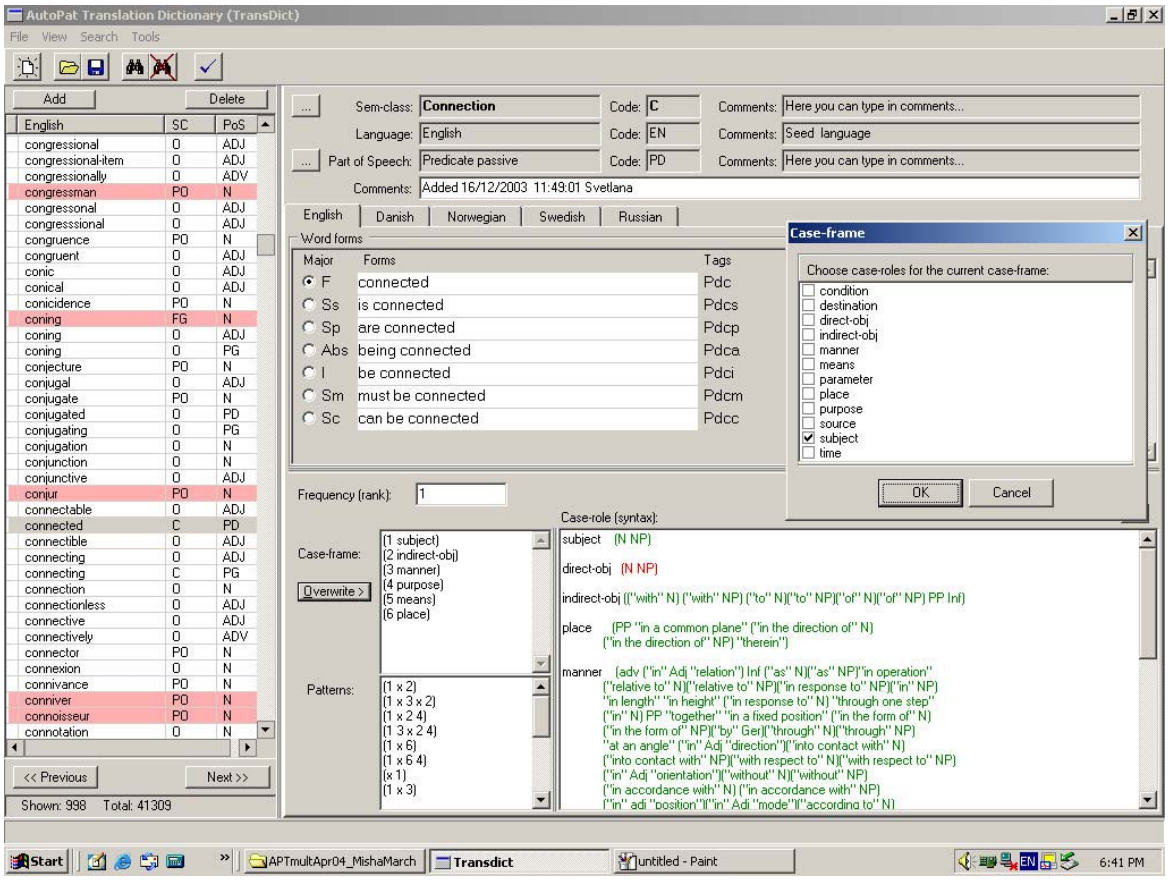

Figure 3. A screenshot of the TransDict interface displaying the entry for the lexeme “connected”.

These extractions contain different data subsets relevant for different processing steps (tagging, disambiguation, transfer and generation).

The extractions are created for every language and for every pair of languages. They are linked to applications by special DLL (dynamic link library) functions that access only one of the dictionary extractions for every processing step. This approach gives a significant increase in access speed and processing, which is crucial for real world systems.

This and the fact that TransDict is implemented for PC motivated our choice not to use the SQL database and XML (which would have slowed down the application performance). It does not mean, however, that TransDict could not be used in the on-line regime. An interface and a DLL can be written for this purpose.

3.3 Supporting tools

We developed the following TransDict tools:

Data importer/merger imports wordlists and/or feature values from external files and applications. For example, the tool is pipelined to a tagger and to application (e.g.,MT) user interfaces, to automatically import unknown words.

Defaulter automatically assignes entry structures and some of feature values to entries.

Editor a) edits feature values in an entry and b) edits dictionary settings, - languages, semantic classes, parts of speech, wordforms and their tags. Any change of settings automatically propagates to corresponding entries.

Morphological generator automatically generates wordforms for a given word base form.

Content and format checker reveals incomplete entries and entries in wrong formats.

Look-up tool performs wild card search and search on any combination of specified parameters.

3.4 Interface design

A lexicographer does not need to use any specification formalizm. A screenshot of the TransDict interface is shown in Fig.2**.** The left pane of the interface screen contains a scrollable list of lexeme base forms[4] in a selected language. Changing the dictionary settings can easily change a base form status of a wordform.

A click on a language bookmark over the morphological zone displays an entry in this language equivalent to a highlighted word in the left column. The main menu contains the selections that switch on the tools.

Figure 3 shows how the default noun entry with two slots for its morphological forms: singular and plural is reset for Danish where definiteness is expressed morphologically, thus duplicating the number of members of the paradigm compared with English.

The "Add" button calls pop-up menus where the developer is prompted to select a semantic class and part-of speech. This done, an entry with a relevant structure, tags and default values will be displayed.

After the user types in a base form all other wordforms are automatically generated on mouse click. The developer is then to review the default knowledge and edit it if necessary. The content and format checker take care of complete and correct descriptions with different kinds of alert messages and rewriting support.

Search can be done either in the look-up or in edit mode.

4 Conclusion

In this paper we concentrated on effort saving in knowledge acquisition for domain tuned MT systems by reducing the size of resources to be acquired, and creating intelligent software for knowledge handling and access speed. We illustrated our approach with a methodology of knowledge specification for patent domain and a tool, a multilingual electronic dictionary, - TransDict, integrated with patent domain applications.

TransDict is an essential part of an on-going project on Machine translation of patent claims. Developing TransDict we focused on such effort saving strategies, knowledge organization, access, reusability, support tools and interface design. As of now (May 2005) the dictionary program including intelligent application adaptive interface integrated with supporting tools and external applications, - AutoPat, AutoTrans, AutoRead (Sheremetyeva, cf.) is fully implemented and tested. This "shell" can now be used to create any number of dictionaries with different feature spaces.

The TransDict patent domain knowledge base currently contains about 80,000 completed English entries and around 300 equivalent Danish entries that are directly used in testing analysis, transfer and generation modules for the English-Danish machine translation system. We plan to increase the English-Danish knowledge base to a product size level by December 2005.

TransDict (with patent domain or other knowledge) can be used as a stand-alone tool, for other applications e.g., for training computational linguists.

[4] For convenience other wordforms are not included in this list but can be displayed on mouse click.

References

Atkins B.T.S. and Levin B. (1991) *Admitting impediments.* In Zernik (ed.) pp.233-62.

Bilac, S and M.Zock. 2003. *Towards a user-friendly dictionary interface.* Papillon 2003 Workshop, 3-5 July, NII, Sapporo, Japan.

Boitet, C., M.Mangeot-Lerebours and G.Sérasset. 2002. *The PAPILLON project: cooperatively building a multilingual lexical data-base to derive open source dictionaries & lexicons.* Proceedings of the 2nd Workshop NLPXML 2002, Post COLING 2002 Workshop. Taipei.

Calzolari N and R. Bindi R. (1990) *Acquisition of lexical information from a large textual Italian corpus.* COLING-90. Proceedings of the 13th International Conference on Computational Linguistics. Helsinki. 3:54-59.Chomsky N. (1995) *The Minimalist Program*. Cambridge, MA:MIT Press.

Church K., Gale W., Hanks P. and D. Hindle D. (1991) *Using statistics in lexical analysis*. In Zernik (ed.), pp.115-64.

Cowie J. and Lehnert W. (1996) *Information Extraction. In Communications of the ACM.* Vol. 39, No.1, January pp.80-91.

Fedder L. 1992, *The Multilex Internal Format*. Multilex report, June.

Fillmore Ch.J. (1970) *Subjects, speakers and roles*. Synthese.21/3/4. pp.251-274.

Hudson R. A. (1995) I*dentifying the linguistic foundations for lexical research and dictionary design.* I Walker at al. (eds), pp.21-51.

Khatchadourian, H. 1992, *Tools, functional specifications*. Multilex report, February.

Onyshkevich, B and S. Nirenburg. 1994. *The lexicon in the scheme of KBMT things*. Thechnical report MCCS-94-277, CRL, NMSU.

Sérasset, G1993. *Recent Trends of Electronic Dictionary Research and Development in Europe*, Technical Memorandum Electronic Dictionary Research (EDR), Tokyo, Japan.

Sheremetyeva, S 2003. Natural Language Analysis of Patent Claims. *Proceedings of the workshop "Patent Corpus Processing" in conjunction with 41st Annual Meeting of the Association for Computational Linguistics (ACL 2003)*, Sapporo. Japan, July 7-12.

K.Wong.2000. Multilingual Electronic Dictionary Project.http://www.csse.monash.edu.au/hons/projects/2000/Kevin.Wong/ksgw.htm