Application Adaptive Electronic Dictionary with Intelligent Interface

Svetlana Sheremetyeva
Copenhagen Business School,
LanA Consulting
Madvigs Alle, 9, 2
Copenhagen, Denmark, DK-1829
lanaconsult@mail.dk

Abstract

The paper presents an electronic dictionary that can be adapted to the needs of different NLP applications. It suggests some ways to save on software customisation and acquisition effort through an intelligent developer interface. The emphasis is made on the flexibility of data representation, handling and access speed.

1 Introduction

In this paper we try to contribute to the problem of electronic dictionaries with a case study, - TransDict, - a multilingual lexicon for a family of patent-related NLP applications, such as AutoPat, APTrans and AutoRead[1]. TransDict thus conforms to the "Multilingual-Specialized" dictionary paradigm (Sérasset, 1993), but it can also be used as a stand alone tool and adapted for other language related tasks, e.g., training computational linguists.

The motivation to focuse on application tuned dictionaries is that though developing reusable full-sized knowledge bases for NLP systems is highly desirable this process is extremely expensive and time consuming, and reusability is not guaranteed. If an NLP system uses a restricted sublanguage, and, thus, can operate with smaller-scale dictionaries, the scope of acquisition and development effort will decrease correspondingly. Dictionary software should be adaptable to the specificity of sublanguages.

The languages that are currently covered are English and Danish but TransDict can easily be extended to a multiple number of other languages. TransDict features a powerful environment for acquisition, editing, browsing, defaulting and coherence checking. It is implemented in C++ as an integral part of 32-bit Windows applications for Windows 95/98/2000/NT.

2 Related work

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 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 is one of the major issues still uderdeveloped (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 Overview of TransDict

3.1 Feature space

TransDict is originally built over a set of features relevant for the patent applications including:

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.

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

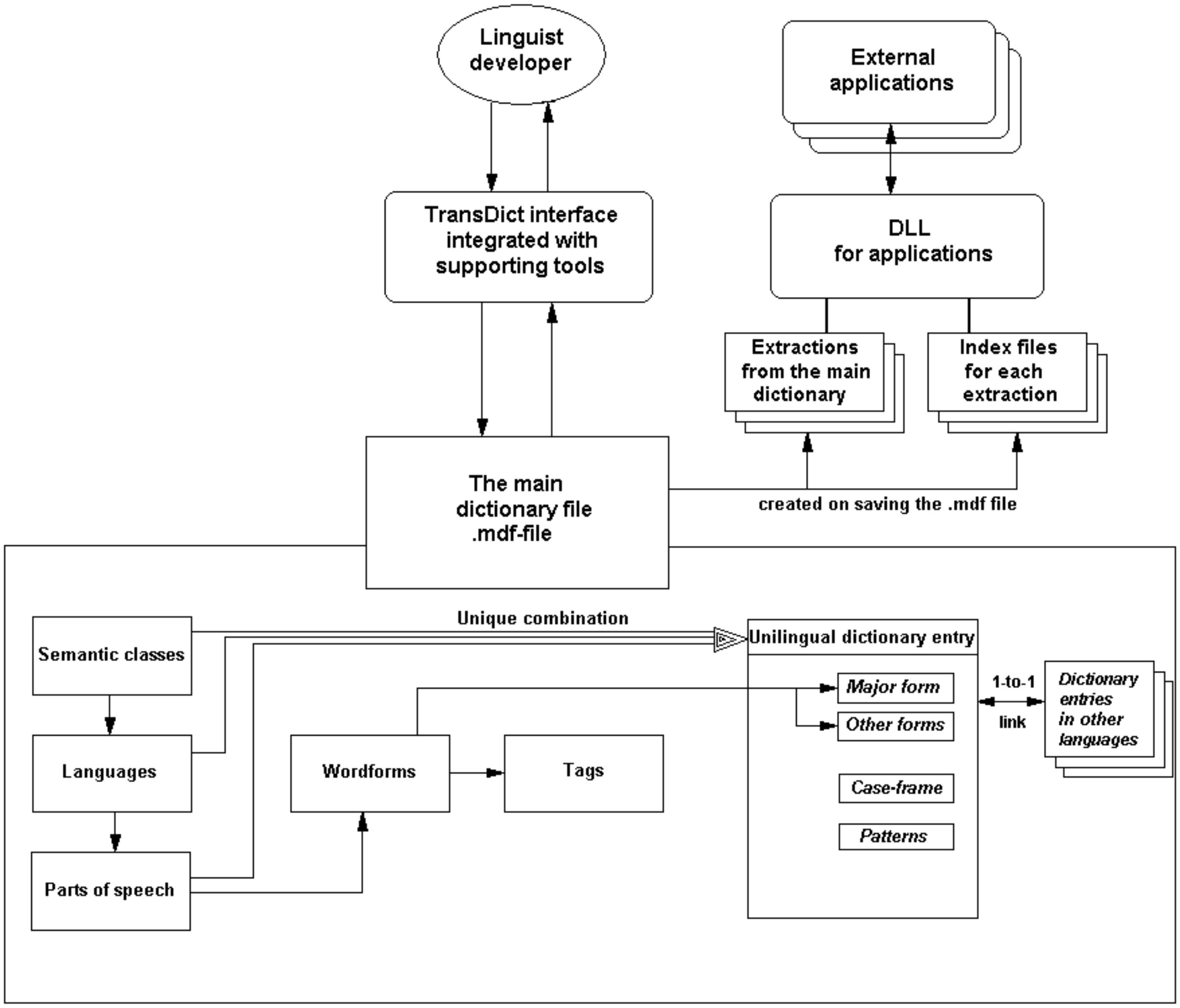

Linguist developer
External applications
TransDict interface integrated with supporting tools
DLL for applications
Extractions from the main dictionary
Index files for each extraction
The main dictionary file .mdf-file
created on saving the .mdf file
Unique combination
Semantic classes
Unilingual dictionary entry
Major form
Other forms
1-to-1 link
Dictionary entries in other languages
Languages
Wordforms
Tags
Case-frame
Patterns
Parts of speech

Figure 1. An overall architecture of TransDict.

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.

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:

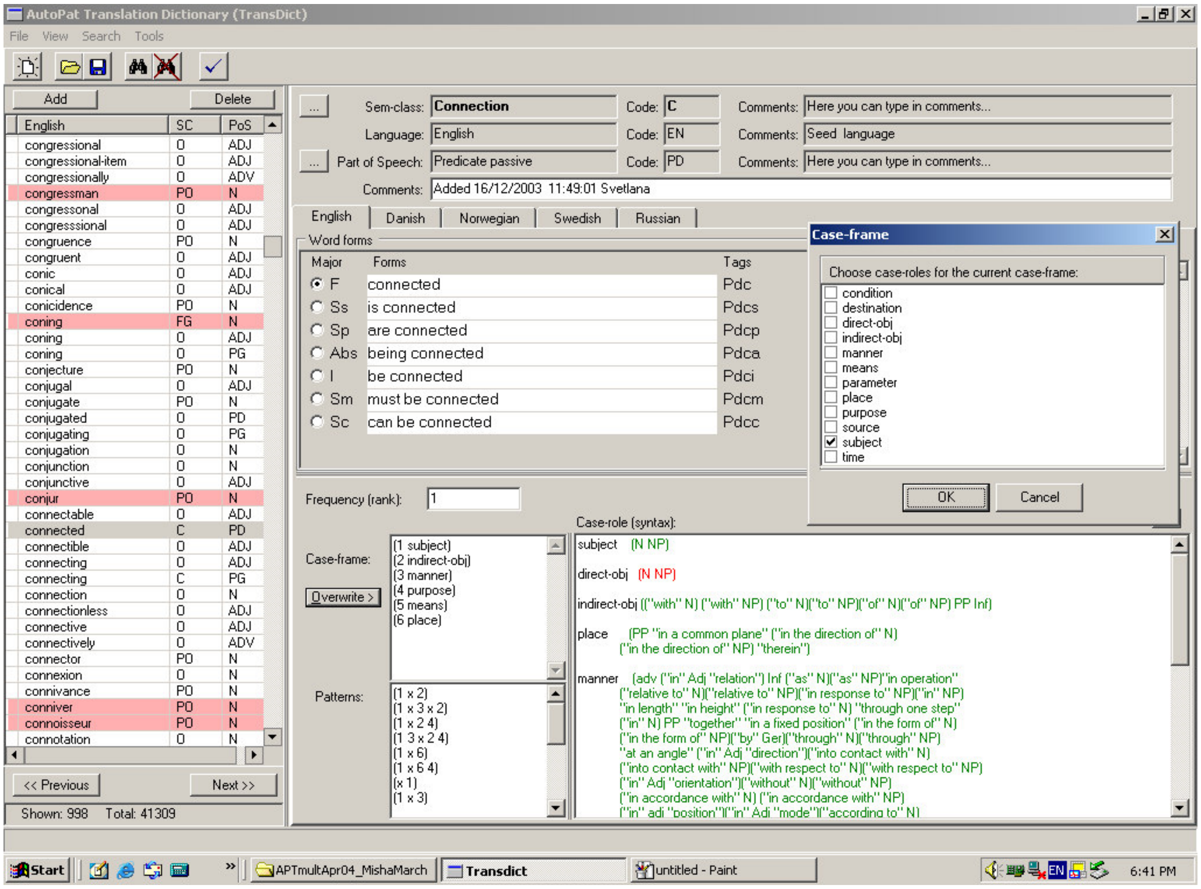
SEM-CL[Language[POS RANK
[MORPH CASE_ROLE FILLER PATTERN]

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 to code several features: POS, number, inflection type and semantic class: object, event, etc., providing for powerful tagging.

The architecture of TransDict is shown in Figure 1. All information is stored in TransDict internal formats: in data files and index files. The developer works with the Main Dictionary File (MDF) visualised by the interface (Figure 2).



English	SC	PoS
congressional	0	ADJ
congressional-item	0	ADJ
congressionally	0	ADV
congressman	P0	N
congressonal	0	ADJ
congresssional	0	ADJ
congruence	P0	N
congruent	0	ADJ
conic	0	ADJ
conical	0	ADJ
conicidence	P0	N
coning	FG	N
coning	0	ADJ
coning	0	PG
conjecture	P0	N
conjugal	0	ADJ
conjugate	P0	N
conjugated	0	PD
conjugating	0	PG
conjugation	0	N
conjunction	0	N
conjunctive	0	ADJ
conjur	P0	N
connectable	0	ADJ
connected	C	PD
connectible	0	ADJ
connecting	0	ADJ
connecting	C	PG
connection	0	N
connectionless	0	ADJ
connective	0	ADJ
connectively	0	ADV
connector	P0	N
connexion	0	N
connivance	P0	N
conniver	P0	N
connoisseur	P0	N
connotation	0	N

Shown: 998 Total: 41309

Sem-class:	Connection	Code:	C	Comments:	Here you can type in comments...
Language:	English	Code:	EN	Comments:	Seed language
Part of Speech:	Predicate passive	Code:	PD	Comments:	Here you can type in comments...

Comments: Added 16/12/2003 11:49:01 Svetlana

Word forms

Major	Forms	Tags
● F	connected	Pdc
○ Ss	is connected	Pdcs
○ Sp	are connected	Pdcp
○ Abs	being connected	Pdca
○ I	be connected	Pdci
○ Sm	must be connected	Pdcm
○ Sc	can be connected	Pdcc

Case-frame

Choose case-roles for the current case-frame:

☐ condition
☐ destination
☐ direct-obj
☐ indirect-obj
☐ manner
☐ means
☐ parameter
☐ place
☐ purpose
☐ source
☑ subject
☐ time

Frequency (rank): 1

Case-frame:
(1 subject)
(2 indirect-obj)
(3 manner)
(4 purpose)
(5 means)
(6 place)

Patterns:
(1 x 2)
(1 x 3 x 2)
(1 x 2 4)
(1 3 x 2 4)
(1 x 6)
(1 x 6 4)
(x 1)
(1 x 3)

Case-role (syntax):

```
subject    (N NP)

direct-obj  (N NP)

indirect-obj (("with" N) ("with" NP) ("to" N)("to" NP)("of" N)("of" NP) PP Inf)

place     (PP "in a common plane" ("in the direction of" N)
          ("in the direction of" NP) "therein")

manner    (adv ("in" Adj "relation") Inf ("as" N)("as" NP)"in operation"
          ("relative to" N)("relative to" NP)("in response to" NP)("in" NP)
          "in length" "in height" ("in response to" N) "through one step"
          ("in" N) PP "together" "in a fixed position" ("in the form of" N)
          ("in the form of" NP)("by" Ger)("through" N)("through" NP)
          "at an angle" ("in" Adj "direction")("into contact with" N)
          ("into contact with" NP)("with respect to" N)("with respect to" NP)
          ("in" Adj "orientation")("without" N)("without" NP)
          ("in accordance with" N) ("in accordance with" NP)
          ("in" adj "position")("in" Adj "mode")("according to" N)
```

Figure 2. A screenshot of the TransDict interface displaying the entry for the lexeme "connected"

When the lexicographer saves the data multiple extractions from MDF are automatically created. 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.

4 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 AutoPat and AutoTrans 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.

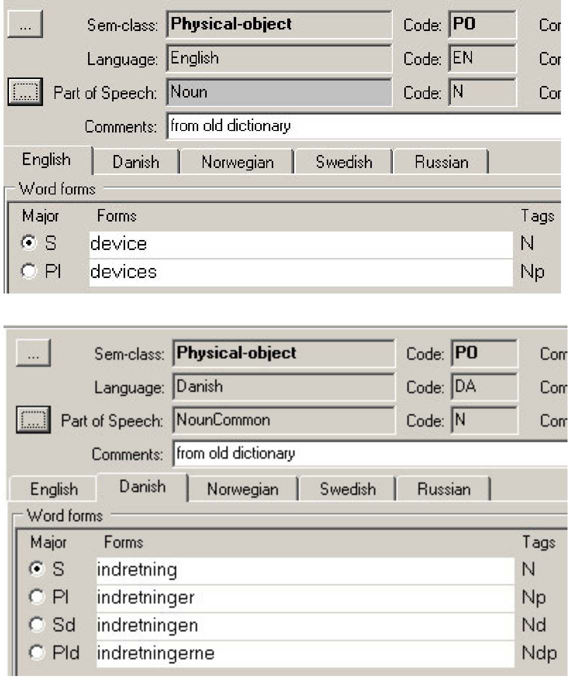
Content and format checker reveals incomplete and/or bad formatted entries.

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

5 Interface design

A lexicographer interacts with the lexicon by an extemely user-friendly interface (Figure2). The left pane of the interface screen contains a scrollable list of lexeme base forms[2] in a selected language. A click on a language bookmark over

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



Sem-class: Physical-object Code: PO
Language: English Code: EN
Part of Speech: Noun Code: N
Comments: from old dictionary

Major	Forms	Tags
● S	device	N
○ Pl	devices	Np

Sem-class: Physical-object Code: PO
Language: Danish Code: DA
Part of Speech: NounCommon Code: N
Comments: from old dictionary

Major	Forms	Tags
● S	indretning	N
○ Pl	indretninger	Np
○ Sd	indretningen	Nd
○ Pld	indretningerne	Ndp

Figure 3. A fragment of English and Danish equivalent entries as shown in the interface.

the morphological zone displays an entry in the selected language equivalent to a highlighted word in the left column. All supporting tools are accessed through the interface menus.

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 to review the default knowledge and edit it if necessary. The content and format checker take care of correct descriptions with different kinds of alert messages and rewriting support. Powerful search can be done both in a look-up and edit mode.

Changing the dictionary settings can easily change a base form status of a wordform, the structure of the entry and other specification parameters. 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 noun paradigm compared with English.

6 Conclusion

In this paper we described an on-going project on developing a multilingual electronic dictionary, - TransDict, integrated with patent domain applications. We focused on such effort saving strategies as knowledge organization, access, reusability, support tools and interface design. As of now (April 2004) 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 60,000 completed English entries and around 100 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 2004.

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

References

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

C.Boitet, 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.

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

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

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

G. Sérasset. 1993. *Recent Trends of Electronic Dictionary Research and Development in Europe*, Technical Memorandum Electronic Dictionary Research (EDR), Tokyo, Japan.

S. Sheremetyeva. 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