DeepDict – A Graphical Corpus-based Dictionary of Word Relations

Eckhard Bick
GrammarSoft & University of Southern Denmark
eckhard.bick@mail.dk

Abstract

In our demonstration, we will present a new type of lexical resource, built from grammatically analysed corpus data. Co-occurrence strength between mother-daughter dependency pairs is used to automatically produce dictionary entries of typical complementation patterns and collocations, in the fashion of an instant monolingual Advanced Learner's dictionary. Entries are supplied to the user in a graphical interface with various thresholds for lexical frequencies as well as absolute and relative co-occurrence frequencies. DeepDict draws its data from Constraint Grammar-analysed corpora, ranging between tens and hundreds of millions of words, covering the major Germanic and Romance languages. Apart from its obvious lexicographical uses, DeepDict also targets teaching environments and translators.

1 Lexicographical motivation

From a lexicographer's point of view, a corpus-based dictionary has a potentially better coverage and legitimacy than a traditional dictionary built on introspection and literature quotes. Many modern dictionaries do therefore make use of corpus data, striving to balance their data with regard to domain, register etc. However, the ultimate product is usually still a traditional dictionary, even in electronic versions, because corpus data are used more for exemplification and simple frequency counts than for dictionary generation proper. Notable exceptions are the *Sketch Engine* (Kilgariff et al. 2004), which uses n-gram collocations and grammatical relations in a systematical way, and the Leipzig University *Wortschatz* project (Biemann et al. 2004), that automatically creates lexical similarity nets from monolingual corpora.

In addition, even where corpora are used selectively or systematically, not all information – especially structural information – is readily accessible, because most corpora of the necessary size will be text corpora without any deeper grammatical annotation. Optimally, the extraction of lexical patterns should not only be based on lemmatized and part-of-speech annotated text, but also exploit true linguistic relations (e.g. subject, object etc.) rather than mere adjacency (n-grams). Finally, even given all of the above, and using a statistics-integrating interface, a lexicographer will only be able to look at one pattern at a time – a tedious process for not least verbs with a complex phrasal and semantic potential. Also, he may not find what he isn't looking for, because the search interface only allows textual searches or because the one resource that might do the job – a syntactic treebank – is usually produced by hand and too small for lexicographical work[1].

The dictionary tool presented here, DeepDict, strives to address both the linguistic quality of available corpus information, and the issue of how to present this information so as to permit a more complete and simultaneous overview of usage patterns for a given word. DeepDict was developed at GrammarSoft and launched commercially at gramtrans.com in September 2007.

2 Ordinary dictionary users

From an ordinary dictionary user's point of view, the following advantages of electronic dictionaries over paper dictionaries should be addressed:

1. There are no size limitations, so the individual entry for an infrequent word can be assigned as much space as for a frequent word, and the exclusion of rare patterns should not be absolute, but governed by user-controlled thresholds.
2. On paper, it is easier to create passive ("definitional") dictionaries than active ("productive-contextual") ones, because the former address native speakers of the target language (TL) , while the latter have to provide a lot of detailed usage information, semantic con-

[1] Size restraints on coverage and statistical salience are mentioned by Kaarel Kaljurand for his *depdict* listings derived from an Estonian treebank, also based on CG, of 100,000 words (http://math.ut.ee/~kaarel/NLP/Programs/Treebank/DepDict/)

Kristiina Jokinen and Eckhard Bick (Eds.)
NODALIDA 2009 Conference Proceedings, pp. 268–271

straints and complementation patterns to a user not familiar with the TL, e.g. A gives x to B (where A, B = +HUM and x,y = -HUM).

3. An electronic dictionary can offer unlimited (linked) corpus examples, on demand, without complicating the entry as such.

3 Assembling the data

Motivated by the arguments discussed in chapters 1 and 2, we opted for Constraint Grammar (Karlsson et al. 1995) as the underlying annotation technique, firstly because of its robustness and good lexical coverage, secondly because its token-based dependency syntax is computationally easier to process. The following method was followed to build the necessary lexico-relational database.

First, for each language, available corpora were annotated with CG parsers and – subsequently – a dependency parser using CG function tags as input (Bick 2005), effectively turning almost a billion words of data into treebanks, with functional dependency links for all words in a sentence[2]. For a number of corpora, only the last step was part of the DeepDict project, since CG annotation had already been performed by the corpus providers for their CorpusEye search interface (http:// corp.hum.sdu.dk). Table 1 provides a rough overview over data set sizes and parsers used.

	Corpus size[3]	Parser[4]	Status[5]
Danish	67+92M mixed	DanGram	+
English	210M mixed	EngGram	+
Esperanto	18N mixed	EspGram	+
French	[67M Wi, Eu]	FrAG	-
German	44M Wi, Eu	GerGram	+
Norwegian	30+20M Wi	Obt / NorGram	+
Portuguese	210M news	PALAVRAS	+
Spanish	50+40M Wi,Eu	HISPAL	+
Swedish	60M news, Eu	SweGram	+

Table 1: Corpora and parsers

In the token-numbered annotation example below, the subject 'Peter' (1. word) and the object 'apples' (6. word) both have dependency-links (#x->y) to the verb 'ate' (2. word).

Peter "**Peter"** <hum> PROP @**SUBJ** #1->2
ate "eat" V IMPF #2->0
a couple of ....
apples "**apple**" <fruit> N P @**ACC** #6->2

From the annotated corpora, dependency pairs ("dep-grams") were harvested – after some filtering between syntactic and semantic head conventions-, using lemma, part of speech and syntactic function. For prepositional phrases both the preposition and its dependent were stored as a unit, de facto treating prepositions like a kind of case marker. For nouns and numerals, in order to prevent an explosion of meaningless lexical complexity, we used category instead of lemma. For nouns, semantic prototypes were stored as a further layer of abstraction (e.g. <hum> and <fruit> in our example). For a verb like 'eat', this would result in dep-grams like the following[6]:

PROP_SUBJ -> eat_V
cat_SUBJ -> eat_V
apple_ACC -> eat_V
mouse_ACC -> eat_V

With little further processing, the result could be represented as a summary "entry" for eat in the following way:

{PROP, cat, <hum>, ...} SUBJ --> eat <-- {apple, mouse, <fruit>, ...} ACC

Obviously, the fields in such an entry would quickly be diluted by the wealth of corpus examples, and one has to distinguish between typical complements and co-occurrences on the one hand, and non-informative "noise" on the other. Therefore, we used a statistical measure for co-occurrence strength[7] to filter out the relevant cases, normalizing the absolute count for a pair a->b against the product of the normal frequencies of a and b in the corpus as a whole:

$$C * \log(p(a\text{->}b)\ ^2 / (p(a) * p(b)))$$

where p() are frequencies and C is a constant introduced to place measures of statistical significance in the single digit range.

[2] Our long-range dependencies provide complete-depth trees, as in constituent treebanks, CG3 dependencies (beta.visl.sdu.dk/constraint_grammar.html) or Functional Dependency Grammar (www.connexor.fi).

[3] Wi = Wikipedia (http://www.wikipedia.com), Eu = the Europarl corpus (Koehn 2005)

[4] More information about the parsers is available at http://beta.visl.sdu.dk/constraint_grammar.html.

[5] The Portuguese, Swedish and Esperanto DeepDicts have unlimited free access, the others have regulated access

[6] Of course, beyond the examples given here, all other relations, such as prepositional objects and adverbials, are equally treated in both the analysis and the interface.

[7] The difference from Church's *Mutual Information* measure is the higher (square) weighting of the actual cooccurrence. This was deemed more supportive of lexicographical purposes – preventing strong but rare or wrong collocations from drowning out common ones.

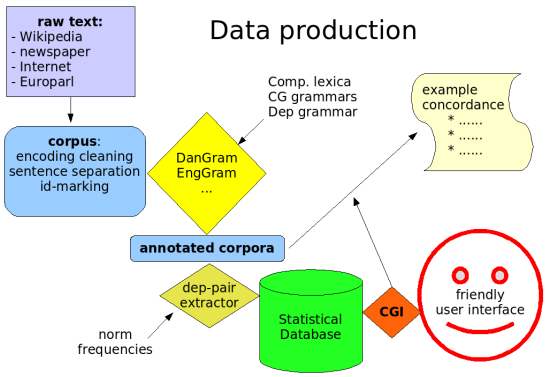

Fig. 1: Data production

The resulting database would then contain, for each dep-gram pair, both its absolute frequency, co-occurrence strength, as well as an index of relevant sentence ID's in the source corpus. Even for a single language, parsing all corpus material and creating the databases, may take days or weeks, and the resulting datasets are so big (currently 90 GB) that querying them in a straightforward fashion would cause unacceptable delays to the user. Hence, special file structures and querying algorithms had to be devised by our interface programmer, Tino Didriksen.

4 The user interface

In order to to meet the requirements outlined in chapter 2, dictionary entries are composed on the fly, respecting user-set significance thresholds[8], and allowing simultaneous overview (a "lexicogram") over a words combinatorial potential. For grammatical reasons, and in order to resolve class ambiguities (e.g. house_N vs. house_V), each word class has its own "lexicogram" template. As can be seen in fig. 2, the lexicogram for the noun 'voice' not only captures typical multiword expressions like "voice actor" and "voice recorder", but also shows typical qualities (loud, deep, husky) and the polysemy implied in "passive voice". The fields of the DeepDict lexicograms are designed to support "natural" reading - which is why the English DeepDict places attributes left and heads right for nouns and adjectives, or subjects left and objects right for verbs, and why other fields are flanked by frame text to create the illusion of a sentence: "one can {recognize, hear, lower, lend, raise} a voice". A minimum of classifier information is provided together with the head word, i.e. gender, transitivity and countability. However, even this information is partly corpus based. Thus, countability/mass is deduced from certain trigger-dependents such as numerals and quantifiers.

voice (noun)

countable

Premodifiers:	PP postmodifiers:	Modifier of:
6.73:7 **loud** · 6.57:7 NUM · 6.41:6 **distinctive** · 5.05:7 **deep** · 6.64:5 **soprano** · 7.46:4 **gravelly** · 7.44:4 **husky** · 4.34:7 **single** · 5.21:6 **inner** · 4.2:7 **own** · 6.59:4 **baritone** · 5.58:5 **passive** · 6.52:4 **hoarse** · 4.46:6 **soft** · 5.46:5 **authoritative** · 4.32:6 **quiet** · 4.28:6 **human** · 6.28:4 **squeaky** · 6.16:4 **narrative** · 5.98:4 **gruff**	8.72:8 **rel-INDP** 1.35:3 interr-INDP 2.58:5 **of character** 2.43:5 **of reason** 2.25:5 **of god** 3.08:4 **from behind** 1.75:5 **of america** 2.7:4 **of conscience** 2.43:3 of dissent	6.04:7 **actor** · 5.94:4 **telephony** · 3.58:5 **actress** · 4.91:3 coil · 2.04:5 **communication** · 2.88:4 **talent** · 2.88:4 **recorder** · 3.52:3 choir · 2.19:4 **transmission** · 1.02:5 **vote** · 1.76:4 **channel** · 2.7:3 characterization · 3.59:2 synthesizer · 3.47:2 inflection · 3.41:2 synthesis · 0.31:5 **system** · 1.27:4 **message** · 1.6:3 directive · 0.51:4 **call** · 1.43:3 lesson

one can ...	14.93:2 modulate · 12.4:2 murmur · 8.62:5 **recognise** · 11.36:1 hush · 10.96:1 shriek · 3.44:8 **hear** · 9.28:2 amplify · 8.34:2 imitate · 3.72:6 **lower** · 6.7:3 obey · 7.2:2 mimic · 4.9:4 **lend** · 5.02:3 possess · 4.68:3 dub · 0.53:7 **raise** · 5.52:2 heed · 4.36:3 drown · 6.14:1 dip · 5.08:2 equal · 6.06:1 sharpen 8.54:4 **creep into** · 10.09:2 exclaim in · 9.09:2 mutter in · 2.63:8 **speak with** · 4.18:5 **sing in** · 8.07:1 retort in · 6.42:2 whisper in · 3.65:4 **reply in** · 4.24:3 cry in · 6.06:1 recite in · 5.78:1 startle by · 4.77:2 inject into · 2.77:4 **listen to** · 0.54:6 **speak in** · 3.39:3 detect in · 3.34:3 consist of · 2.91:3 shout in · 2.04:3 sing with · 0.78:3 sound like	a voice
a voice can ...	14.14:3 muffle · 12.44:4 **tremble** · 9.54:4 **whisper** · 11.44:2 crackle · 11.32:2 growl · 6.13:7 **sound** · 10.61:1 wobble · 9.49:2 drip · 9.49:2 thicken · 10.29:1 squeak · 7.85:3 falter · 5.82:5 **echo** · 8.69:2 waver · 7.43:3 harden · 8.24:2 reverberate · 8.7:1 exclaim · 5.38:4 **fade** · 5.25:4 **shout** · 6.17:3 deepen · 7.17:2 startle	
a voice can be	13.33:3 muffle · 12.39:2 hush · 8.1:3 dip · 9.75:1 tinge · 8.7:1 amplify · 1.58:7 **hear** · 4.68:3 dub · 5.12:2 choke · 4.09:2 drown · 3.62:2 strain	...'ed

Fig. 2: DeepDict noun template

The co-occurrence strength between the lookup word and a given relation is presented in red numbers in front of the context word, separated by a colon from the absolute frequency class (an integer representing the dual logarithm of the actual frequency[9]. Ordering is a function of these 2 values, and to give further salience to important correlations, frequency classes of 4 and above are in bold face. At the same time, the red numbers serve as clickable links to a corpus concordance for the relation in question – allowing lexicographers to check DeepDict's analysis in rare or problematic cases, especially if low significance thresholds have been set by the user.

Personal and quantifier pronouns are so frequent that exact statistical measures are of little interest. However, they may provide semantic information in a prototypical fashion, and they are therefore listed - by order of frequency - at the top of subject and object fields. Personal pronouns may help classify activities as typically male (he) or female (she), or mark objects as inanimate (it) or mass nouns (much). Even sociolinguistic deductions are possible: Thus the DeepDict entry for the verb "caress" (Fig. 3) shows, that males (he) are more likely to caress females (she) than vice versa.

[8] There are 4 types of threshold: (a) minimum occurrence, designed to filter out corpus errors and hapaxes, (b) minimum co-occurrence strength, with a default at 0, (c) maximum number of hits shown per field, and (d) minimum lexical frequency of relation words, for language learners, so rare words will be explained with ordinary word contexts rather than vice versa.

[9] In its default settings, the interface cuts out relations with frequencies < 4, to avoid errors caused by misspellings and other corpus anomalies, or faulty analysis.

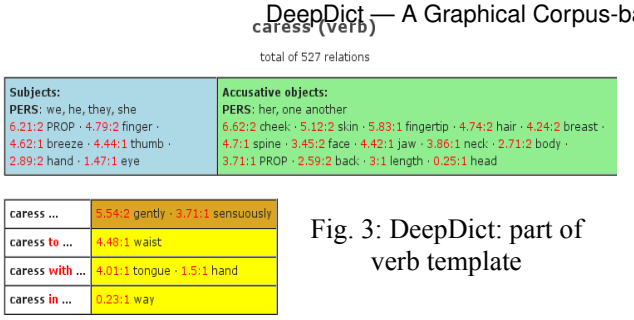

Fig. 3: DeepDict: part of verb template

The example also illustrates metaphorical usage – the lexicogram not only lists the bodyparts that do the caressing (subject) and the ones that are caressed (objects), but also mentions 'eyes' and even 'breeze' as caressors. Finally, it shows how prepositions (*with tongue/hand*) are linked into the verb template. For other verbs, it is here we will find prepositional valency, too.

Adverb-verb collocations may appear in several functional shades, ranging from (a) free temporal, locative and modal adverbs (work *where/when/how*) to (b) valency bound adverbial complements (feel *how,* go *where*) and (c) verb-integrated particles (give *up,* fall *apart*). In some cases, it may even be difficult to decide on one or other category (eat *out*). Since DeepDict is basically intended as a dictionary tool, syntactic hair splitting is less important, and only the verb particles (c) are singled out, to cover phrasal verbs, presenting the rest in a single (brown) field ('gently/sensuously' for the verb 'caress').

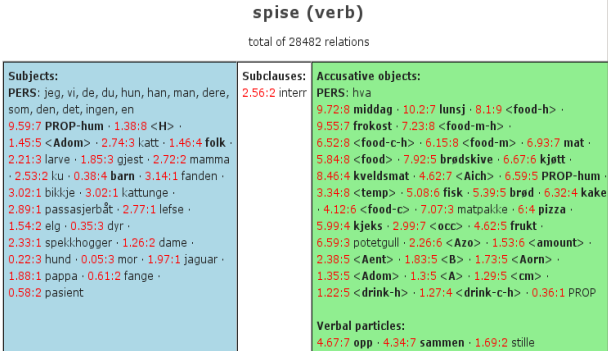

Fig. 4: Semantic prototypes

In the parsers providing the corpus data behind DeepDict, nouns are classified according to semantic prototype class[10], e.g. as <Hprof> (professional human) or <tool-cut> (cutting tool) or <Vair> (air vehicle), and this semantic generalisation has been made available for some DeepDict languages. In the conference demo linked to this paper DeepDict will be accessible through an internet portal at (http://www.gramtrans.com).

5 Conclusion and future work

We have shown how syntactically related word pairs can be harvested from Constraint Grammar-annotated dependency corpora and fed into a statistical database that will allow the on-the-fly creation of so-called "DeepDict lexicograms" – semi-graphical overview pages for dictionary words, with information about head and modifier selection restrictions, verb complementation and phrasal collocations. The tool allows lexicographers to mine corpora not only for examples of structures and lexical relations, but for the structures and relations themselves. DeepDict can be chained to other lexical resources - traditional definition dictionaries, ontologies or bilingual dictionaries (cp. the QuickDict dictionaries at gramtrans.com). Since the DeepDict method can be run from scratch on any language data accessible to a CG parser, it should be possible in the future to provide researchers, lexicographers and teachers with individual DeepDict installments for specific user corpora, reflecting a specific domain, genre or language variety.

References

Bick, Eckhard. (2005) "Turning Constraint Grammar Data into Running Dependency Treebanks". In: Civit, Montserrat & Kübler, Sandra & Martí, Ma. Antònia (red.), *Proceedings of TLT 2005*, Barcelona, December 9th - 10th, 2005), pp.19-27

Bick, Eckhard (2006): "A Constraint Grammar-Based Parser for Spanish", *Proceedings of TIL 2006 - 4th Workshop on Information and HLT.*

Biemann, Chris & Stefan Bordag & Uwe Quasthoff & Christian Wolff (2004). "Language-Independent Methods for Compiling Monolingual Lexical Data". In *Comp. Linguistics and Intelligent Text Processing.* Springer: Berlin, pp. 217-228

Church, Ken and P. Hanks 1991. Word Association Norms, Mutual Information and Lexicography. Computational Linguistics,Vol.16:1, pp. 22-29.

Karlsson, Fred et al. (1995): *Constraint Grammar - A Language-Independent System for Parsing Unrestricted Text.* Natural Language Processing, No 4. Berlin & New York: Mouton de Gruyter.

Kilgarriff, Adam, Rychlý, P., Smrž, P. & Tugwell, D. (2004). "The Sketch Engine". Paper presented at EURALEX, Lorient, France, July 2004.

Koehn, Philipp (2005). Europarl: A Multilingual Corpus for the Evaluation of Machine Translation. MT Summit X, Sept.12-16, 2005.Phuket,Thailand.

[10] Depending on the language, about 160-200 prototypes are used (http://beta.visl.sdu.dk/semantic_prototypes_overview.pdf). For our purposes, semantic prototypes were preferred to classical wordnets because the latter have too many (and sometimes usage-dependent) subdistinctions and do not clearly state where in a hyperonomy chain to find the best classifier.

ISSN 1736-6305 Vol. 4
http://hdl.handle.net/10062/9206