Towards Automatic Translation of Support Verbs Constructions:

the Case of Polish *robić/zrobić* and Swedish *göra*

Elżbieta Dura, Barbara Gawrońska

Lexware Labs, Högskolan i Skövde, School of Humanities and Informatics

elzbieta@lexwarelabs.com barbara.gawronska@his.se

Abstract

Support verb constructions range from idiosyncratic to predictable. Lexical functions provide a solution to translation of idiosyncratic constructions only. Our corpus research aims to contribute to automatic translation of support verb constructions where the verb selects certain semantic groups of collocates, and where novel collocations can be expected. We investigate samples of support verb constructions with Polish *robić/zrobić* and Swedish *göra*. Nouns attested on the Internet as objects of these verbs are subdivided into semantic groups. Translation rules are then proposed for each group, and the similarities and differences in the behaviour of the verbs in both languages are discussed.

Introduction

Most linguistic theories focus on paradigmatic relations between words. The shift of attention towards syntagmatic relations was advocated by Firth (1957, 1968), who proposed the term "collocation". However, Firth's notion of collocation is a very general one, and many linguists after him (Mitchel 1971, Cowie 1981, Hausmann 1985, Kjellmer 1987, Sinclair 1987, Moon 1998 - this list is far from complete) tried to formulate more stringent definitions in order to distinguish between collocations, idioms, compounds, and regular syntactic constructions. It is out of scope of this presentation to discuss the different definitions in detail. For the purpose of further discussion, we adopt the distinction between regular syntactic phrases, collocations, and idioms, suggested by Mel'čuk at Euralex 1990 (as related by Heid 1994:233). The discriminating factor in this classification is the degree of compositionality:

- In idioms, none of the components contributes to the semantics of the phrase.
- In collocations, one of the components contributes to the semantics of the phrase.
- In regular syntactic phrases, all components contribute to the semantics of the phrase.

Knowledge of collocations is extremely important for human language learners, human translators, and for NLP-applications, such as Machine Translation or Information Extraction. Despite this, the current representation of phrases with support verbs in electronic dictionaries is far from sufficient. During the last two decades, the growing access to electronic corpora, and, consequently, the rapid development of corpus linguistics, has had a great impact on the work on identification and extraction of multiword entries (Church et al. 1991). However, as pointed out by Danielsson (2001:35), "in most collocation studies in computational linguistics, the focus is more on what can be retrieved automatically from large corpora than on what role the results might play in the language".

The main problem is not the lack of language-specific collocation dictionaries (Mel'čuk et al. (1984, 1988), Benson et al. (1986), Kjellmer (1994)), but the shortage of multilingual resources, where collocations would be linked to their translation equivalents in a way that would be linguistically consistent and easily accessible for both human users and Natural Language Processing systems. For example, one of the shortcomings of the (in many ways very useful) large lexical database EuroWordNet (Fellbaum 1998, Vossen 1998, Viberg et al. 2002) is the lack of collocational links between nouns and verbs, something that makes it difficult to distinguish between the 'contentful' uses of a verb and its support function in automatic text processing. An attempt to overcome this problem is currently made within the Berkeley FrameNet Project (Baker et al. 1998, Gildea and Jurafsky 2002), which aims at a lexical database for English containing a detailed description of the syntactic and semantic valence.

The work presented here may be regarded as a step towards a better representation of collocational links for the purpose of NLP. The apparatus of lexical functions works only for support verb constructions entered in a dictionary but not for novel constructions, which are focused here.

Support verb constructions (svc:s)

Many collocations display the pattern V + N and have one-word synonyms, e.g. *make a decision* - *decide*, *make a discovery* - *discover*. These constructions are even called "dissolved verbs", since they function as verbs (predicators) in a sentence. The grammatical verb in these constructions has often a very general, "light" semantics and supports the semantically "heavy" direct object NP. Verbs occurring in this function have been called "delexical verbs" (Collins Cobuild 1992), "vacuous auxiliaries" (Wilks et al. 1996) and "support verbs" (Heid 1994). In the following, we will use the term "support verbs". The nouns occurring as objects of support verbs are usually non-referential. As pointed out by Fillmore (2003), the nominal object in svc:s "cannot really be interrogated": a question-answer pair like: *What have you*

made? – A decision to go home is not a natural conversation.

Phrases with support verbs are very frequent and pose serious problems for Machine Translation, since, in many cases, the support verb should not be translated by the default equivalent of the "heavy", sense of the verb. It is for example incorrect to translate *take a break* into Polish as **wziąć przerwę* or *make a speech* as **robić przemowienie.*

Most collocations with support verbs are not completely idiosyncratic. A lot of them belong to the type called by Martin (1992) and Heid (1994) "conceptual collocations". In a conceptual collocation, a co-occurent (e.g. a support verb) does not combine with one single term only; instead, it selects a group of terms that normally share certain semantic features. Conceptual collocations can therefore be described in terms of selectional restrictions, and this in turn may enable a direct transfer of svc components in machine translation.

The goal

The goal of the intended project as a whole is to investigate the repertoire of support verb constructions denoting acts of communication and/or cognition in Polish, English, and Swedish (like *make a complaint/declaration/remark*, *deliver a lecture*, *issue a denial*) for the purpose of machine translation. The most frequent English support verb in this domain is *make.* It is used as a kind of a joker support verb, even in contexts for which there may be specific lexicalized variants, e.g. *deliver a speech* and *make a speech*. It would be advantageous for a machine translation system to use such joker verbs in cases when it lacks the whole phrase in its lexicon. Such graceful degradation of a machine translation system can be achieved if the system is provided with information about the context in which it is possible to use a joker verb in a given language. Direct lexical transfer can thus be enabled with support verbs represented as independent lexical entries. The investigation presented here is meant as a step towards creation of such lexical structures.

The default translation equivalents of *make* in its basic (concrete) sense are Swedish *göra* and Polish *robić/zrobić*. These verbs, however, behave differently when used in svc:s. The part reported in the present paper focuses on the comparison between *göra* and *robić/zrobić* as support verbs.

The methodology of the corpus research

A prerequisite for an adequate lexical representation should be corpus research followed by linguistic analysis. The methodology of corpus research employed here follows L'Homme and Bertrand (2000) and consists of:

1. Selection of preliminary key words.
2. Extraction of verb-noun combinations in which the key words are used
3. Selection of co-occurents (support verbs) and new extraction of combinations with the selected co-occurents
4. Analysis, classification and description of the extracted combinations.

The preliminary key words (step 1) were nouns classified as "statements" (*speech, lecture, complaint*…) in FrameNet. We continued by an investigation of verb-noun collocations (steps 3 and 4). Here, we present the part of the analysis and classification work which concerns Polish *robić/zrobić* and Swedish *göra*.[1]

Extraction from the Internet

Extractions from the Internet were performed with Lexware Culler (Dura 2004) - a concordancer mounted on Google. Google ranking is determined mainly by the number of links to a webpage, and it does not provide more than 2 excerpts per website, which luckily for a corpus linguist excludes very odd language uses. At the same time, excerpts are obtained from a variety of sources.

We found 3 810 000 webpages with at least one occurrence of the verb *robić* (40 inflectional forms) and 3 780 000 of the verb *zrobić* (38 inflectional forms) not followed by the reflexive pronoun *się/sobie*. 7 570 000 Swedish webpages had at least one occurrence of *göra* (4 inflectional forms) not followed by the reflexive *sig*.

The following search phrases were entered to Culler: for Polish *robić &noun*, *zrobić &noun*, and for Swedish *göra* followed by a bare noun, *göra* followed by an indefinite article: uter *en* or neuter *ett*.[2] The quantitative results of the extraction from the Internet samples are shown in Table 1.

Search phrase	Total excerpts	Noun tokens	Noun types	Noun lexemes
göra &noun	987	225	162	77
göra en &noun	576	296	211	146
göra ett &noun	580	286	171	125
robić &noun	3 191	543	289	135
zrobić &noun	3 219	439	224	91

Tab.1 The quantitative results of the extraction

The preselected excerpts were examined manually. Excerpts displaying other government patterns than V NP were eliminated, for example V NP ADJ, such as *robić sytuację nieznośną* (*to make the situation unbearable*). Instances of basic sense uses, i.e. phrases referring to the production of concrete physical objects, were also removed. The remaining material contained 135 direct object lexical nouns for *robić* , 91 for *zrobić*, and 348 for *göra*.

The analysis

The following semantic classes of svc:s could be distinguished in the obtained excerpts:

[1] In the following text the English verb *make* is used in all translations, apart from an example in which aspectual differences are relevant.

[2] &noun is a variable interpreted by Culler as any noun in any form in the marked position, here following the verb.

1. NarrArt: produce a narrative artefact: *translation*, *report*, *comments*, *musical, video, documentary*,
2. GoalAct: perform a goal-oriented activity: *research*, *career*, *business*, *education*, *trip*,
3. BodyMan.: perform bodily manipulation: *manicure*, *nails*, *abortion*, *operation*, *hair*,
4. BodyMov: perform bodily movement: *gestures*, *steps*, *salto*, *grimaces*,
5. SocEv: organize or participate in a social event: *revolution*, *mascarade*, *party*, *conference*, *meeting*,
6. ExcBeh: evoke by one's behaviour a state perceived as unusual/exceptional: *noise*, *confusion*, *hell on earth*, *impression, wonders*,
7. Other: be engaged in an activity: *exceptions*, *obstacles*, *success*, *difference*, *self-examination*, *exchange.*

Table 2 shows the distribution of constructions in the distinguished classes. The seventh group, labeled as "other" above constitutes 16% of the Polish and 20% of the Swedish material. Some of these miscellaneous constructions constitute very salient sub-groups of nouns but these are not individuated here because of an insufficient number of instances. One such group is to form some shape of components, such as *make a row, a circle*, which appears in both languages. Yet another group is analogous to performing manipulation. It comprises objects which need reparation, with similar examples in both languages, e.g. *make a tire*.

The two last columns in Table 2 show how many of the constructions in a specified group can be translated directly (word-for-word) from Polish into Swedish and from Swedish into Polish.

	Group	Pl	Sw	Pl →Sw	Se → Pl
1	Narr.Art.	28%	33%	88%	89%
2	Goal Act.	14%	34%	89%	52%
3	Body Man.	7%	3%	50%	20%
4	Body Mov.	5%	2%	80%	60%
5	Soc. Ev.	14%	2%	43%	80%
6	Exc. Beh.	14%	6%	57%	100%

Tab. 2 The distribution of the distinguished classes and the percentage of possible word-for-word translations

One of the chief purposes of using svc:s instead of their simple verb counterparts (when such are available) is the shift of focus, e.g. *robić dodawanie* vs *dodawać* (*make additions*), *göra beräkningar* vs *beräkna.* 13% of the Polish constructions and 11% of the Swedish constructions are cases of this kind of verb expansion where a simple verb counterpart is available in the lexicon. The quantitative difference is not big and it is mainly due to excessive compounding in Swedish, e.g. *göra besök* (*make a visit*) is simply *besöka* while *göra studiebesök* (*make a study visit*) does not have a one word verb counterpart.

Besides this outstanding difference in compounding it rather seems to be a matter of coincidence whether some situation types have one verb lexicalizations in a language, e.g. one can say in Polish *dziwnie gestykulować* (*to gesture strangely*) instead of *robić dziwne gesty* (*make strange gestures*) but there is no **dziwnie minować*[3] for *robić dziwne miny* (*make strange faces*).

Produce narrative artefact - NarrArt

NarrArt encompasses svc:s with nouns primarily denoting narrative artefacts, including so-called "picture nouns" (*photograph, picture*). *Robić*/*zrobić* and *göra* are used here in one of their basic senses: "produce/create". Not all constructions in group 1 classify as svc:s. Some of them are instances of systemic polysemy of the full verb *robić*/*zrobić* and *göra*. It is nonetheless relevant to investigate whether and when the verbs in the two languages can be used as translation equivalents.

As shown in Table 2, the majority of group 1 constructions may be translated directly between Polish and Swedish. However, Polish seems to put stronger restrictions upon what can be regarded as a narrative artefact. *Robić/zrobić* normally does not combine with nouns that primarily denote oral communication acts (even if their meaning may be extended to include written messages), like *statement, utterance*. Swedish *göra* collocates frequently with these nouns.

An interesting detail is that the only nouns that can refer to oral communication and occur with *robić/zrobić* in our material are *comments* and *remarks*. Both share the semantic feature of adding something to an already existing utterance/narrative artefact. The nouns that may refer either to oral communication or to written artefacts require further subcategorization for the purpose of translation, while in the case of visual narrative artefacts (movies, pictures etc.) direct transfer may apply.

Direct transfer from Polish into Swedish is not possible in constructions referring to events on the wedge between a narrative artefact and social event, like *czat* (*chatting*). This is probably due to the generally low frequency of *göra* in constructions denoting social events (see Table 2).

Perform a goal-oriented activity - GoalAct

NarrArt focuses on the product, while in GoalAct the activity is in focus. It is not always easy to draw a borderline between the two groups. For instance, *translation* may be viewed either as a goal oriented process, or as its result: the text. As a criterion for distinguishing between artefact and activity we used the following test: if a noun X can occur in a question like: *where is your X?,* it is to be regarded as denoting an artefact. *Translation, report, comment* were therefore classified as NarrArt. Nouns that cannot occur in this context (if the question is not meant to be ironic) and that are possible in a context like: *How long are you going to continue (with) your X* (*X = research, career, business…*) were regarded as GoalAct.

The overlap in word-for-word translation from Swedish and Polish is not high here, while almost 90% of the Polish constructions can be translated directly. Swedish *göra* appears to have more general sense than Polish *robić/zrobić*. It can be used in contexts such as *ärende* (*errand*), *arbete* (*work*, *job*), *handledning* (*supervision*).

The group of GoalAct has a significant overlap in translation from Swedish into Polish, but there is an

[3] At least not in standard Polish.

important exception. There is a systemic polysemy in Polish which is absent in Swedish: an educational organization in Polish can also mean a degree/diploma obtained from it. For instance, *zrobić uniwersytet* means to get a degree from a university.

Perform bodily manipulation - BodyMan

Constructions of BodyMan constitute a well pronounced distinct group.

There is an important discrepancy between Swedish and Polish in this group. A default interpretation of the subject in Swedish is patient (the semantic object of the manipulation) while it is agent in Polish. For instance, in Swedish, the subject of *göra abort* refers to the one who undergoes an abortion, while the default interpretation of the subject in Polish is the person who performs the abortion. The second interpretation (subject=patient) is not excluded, but it requires a context that makes the disambiguation possible. The subject of e.g. *robić*/*zrobić paznokcie* may refer to the manicurist or the customer. An unambiguous reading of the subject as patient appears when the dative pronoun *sobie* is added. The low overlap between Swedish and Polish in this category is due to this difference in default interpretation of argument roles.

In both languages an abstract sense of a concrete noun is coerced in svc:s of this type. Nouns occurring in svc:s are normally abstract deverbal nouns, but in the BodyMan group we find a relatively high amount of concrete nouns, like *hair*, *nails*, *eyebrows*. For instance, *robić*/*zrobić paznokcie* (*do nails*) is equivalent to *robić*/*zrobić manicure* (*do manicure*). The noun *paznokcie* in this context does not refer to concrete body parts; instead, it is used to specify the manipulation act. In both languages the sense of bodily manipulation is evoked for objects of manipulation such as *nails*, *eyebrows* (in the sense of doing hair or eyebrows in a beauty parlor).

Perform bodily movement - BodyMov

The overlap between Swedish and Polish in this group is 60%. The discrepancies are mainly due to the fact that many natural movements, such as walking or breathing, are not "made", but "taken" in Swedish: *ta ett steg* (*take a step*), *ta ett andetag* (*take a breath*). In Polish it is perfectly natural to *make a breath* (*robić wdech*) or *make an exhalation* (*robić wydech*). Similar metaphorical extensions can be noted here in both languages, e.g. *make a move* can occur in an abstract sense in contexts such as *on a chess board*, *in one's career*, etc.

Organize or participate in a social event – SocEv

Constructions in this group display 80% overlap in translation from Swedish into Polish, while the translation of Polish construction in our materinal into Swedish is possible only in 43% of cases. It is worth noting that this category is quite frequent in Polish (14% of all phrases in our material) and rare in Swedish (2%). A significant difference between Swedish and Polish is the fact that the interpretation "be engaged in an organized activity" is the preferred one in Swedish. Swedish *göra revolt* (*make a revolt*), means rather *to participate in a revolt* than *to organize a revolt*. *Robić/zrobić marsz pokoju* means *to organize a peace march* while *göra en fredsmarsch* means *to go on a peace march*. The safest translation of *robić/zrobić* from Polish into Swedish in these contexts is thus *ordna* or *organisera* (*organize*).

Evoke a state perceived as unusual - ExcBeh

The class of ExcBeh is well represented in the material of both languages. It encompasses constructions denoting states of affairs of which the perceiver is a necessary element. The situation is judged as different from what is normal or neutral (in a positive or negative sense): *mistake, confusion, hell on earth, wonders*.

The overlap in translation from Swedish and Polish is total. It is not the case when translating from Polish into Swedish. Constructions with e.g. *oväsen* and *hałas* (*noise*), *spratt* and *kawały* (*jokes*) can be translated directly, but there are contexts which require an explicit statement of creation in Swedish. For instance, confusion cannot simply be "made" in Swedish; the right expression is *skapa förvirring* (*create confusion*) while *robić/zrobić* is perfectly suitable in Polish in the corresponding context: *robić zamęt*/*zamieszanie*.

It needs to be noted here that register plays an important role in svc:s, particularly when light support verbs are involved. For instance *zamęt* (*confusion*, *muddle*) is a more colloquial noun than *förvirring* (*confusion*), and it is possible to use it with *robić*/*zrobić*, while *göra* is not suitable for *förvirring, skapa* (*create*) needs to be used instead.

Hurdles on the way to direct translation

The two realms in which the two languages differ significantly is compounding in Swedish and aspect in Polish. These differences particularly limit the possibility of the use of direct transfer in automatic translation.

In Swedish the modifying component of a compound may change the character of the whole complement and thus the whole svc. For instance, a direct counterpart of *göra en bedömning* (*make a judgement*) is *robić ocenę sytuacji* while *göra en felbedömning* (*make a misjudgement*) requires a different construction in Polish e.g. *pomylić się w ocenie sytuacji*.

The category of aspect has not been taken up here, not because it is not significant, but because it would require separate study (cf Jędrzejko 1998). Aspect seems to be responsible for negation appearing in some Swedish equivalents of Polish expressions. Despite the same use of the noun *shit* in svc:s of the two languages an affirmative is not possible in Swedish. The equivalent of *robić gówno* (lit. *make shit,* meaning*: "*do nothing") is *inte göra ett skit* (*not to do a shit*). An affirmative Swedish construction *göra ett skit* means *to produce a shit*.

Aspect seems to be involved in a number of differences in the analyzed material. For instance, verbs which focus on a quick accomplishment frequently correspond to a Polish perfective verb, e.g. Swedish *tåga in* and Polish *wkroczyć* (*march in*). In order to express durativity or iterativity an svc is required in Swedish: *göra intåg* (lit. *make in-march*) when an imperfective verb is used in such contexts in Polish: *wkraczać*.

Conclusions

Swedish *göra* occurs only marginally with nouns denoting social events, while objects belonging to this

category are frequent in Polish. If the object refers to a social event, the whole construction means 'participate' in Swedish, rather than 'organize' or 'create/evoke', which is the default reading of svc:s in this subgroup in Polish. A machine translation system would benefit from a rule changing the Polish support verb into Swedsih *ordna* or *organisera* in cases when the object belongs to the category "social events".

The Polish verbs *robić/zrobić* share their basic, semantically heavy sense: "produce, create" with the Swedish verb *göra:* When used as support verbs, they also display considerable overlap, which, if noted properly in the dictionaries, can be used for direct translation. Swedish *göra* combines with a broader range of nouns that denote narrative artefacts and goal oriented activities than Polish *robić/zrobić*. This difference is probably due to the absence of aspect in Swedish (particularly durative). The role of the category of aspect, as well as the role of register, require further investigation.

Considering the fact that novel svc:s often appear in both languages, a compositional account of svc:s is a prerequisite for successful translation. Our investigation confirms both the need to distinguish classes of nouns co-occurring with support verbs and the need to separate support verbs from the corresponding full verbs.

References

Baker, C. F., Ch. J. Fillmore, and J. B. Lowe. 1998. "The Berkeley FrameNet project." In:*Proceedings of COLING/ACL-98.* Montreal.

Benson, M., E. Benson and R. Ilson. 1986. *The BBI combinatory dictionary of English: a guide to word combinations.* Philadelphia: John Benjamins Publishing Company.

Church, K., Gale, W., Hanks, P., and Hindle, D. 1991. "Using Statistics in Lexical Analysis". In: Zernik, U. (ed.) *Lexical Acquisition: Exploiting On-Line Resources to Build a Lexicon,* L. Erlbaum Associates, Hillsdale.

Collins Cobuild. 1992. *English Usage*. London: HarperCollins Publishers.

Cowie, A. P. 1981. "The treatment of collocations and idioms in learner's dictionaries."In: *Applied Linguistics,* 2(3).

Danielsson, P. 2001. *The Automatic Identification of Meaningful Units in Language.* PhD Thesis, Språkdata, Dept. of Swedish, Göteborg University.

Dura, E. 2004. "Concordances of Snippets". *Coling Workshop on Using and Enhancing Electronic Dictionaries*. Geneva.

Fellbaum, C. (ed.) 1998. *WordNet : An electronic lexical database.* Cambridge (Mass.): The MIT Press.

Fillmore, Ch. 2003. "Multiword Expressions: An Extremist Approach". A lecture delivered at the conference in Berlin (Magnus-Haus) September 18-20 *Collocations and idioms: linguistic, computational, and psycholinguistic perspectives*. Available at url: www.bbaw.de/forschung/kollokationen/documents/coll_fillmore_mwe.pdf

Firth, J.R. 1957. "Modes of Meaning". In: *Papers in Linguistics 1934-1951*, pp. 190-215, Oxford University Press, Oxford.

Firth, J.R. 1968. "A synopsis of lingustic thoery". In: Palmer, F.R. (ed.) *Selected Papers of J. R: Firth 1952-1959.*

Gildea, D. and D. Jurafsky. 2002. "Automatic labeling of semantic roles." In: *Computational Linguistics,* 28(3).

Hausmann, F. J. 1985. "Kollokationen im deutschen Wörterbuch". In: H. Bergenholtz & J. Mugdan (eds), *Akten des Essener Kolloquiums zur Grammatik im Wörterbuch.* Tübingen: Lexicographica: Series Maior, 3.

Heid, U. 1994. "On ways words work together: research topics in lexical combinatorics." In: *Proceedings of Euralex-94 International Congress.*

Jędrzejko, E. 1998. "Stary problem, nowe możliwości. Uwagi o analityczności w słowniku i wariancji mechanizmów znakotwórczych". In: Jędrzejko, E. (ed.) *Nowe czasy, nowe języki, nowe i (stare) problemy*.

Kjellmer, G. 1987. "Aspects of English Collocations", In: Mijs, W. (ed.) *Corpus Linguistics and Beyond,* pp. 133-140, Rodopi, Amsterdam.

Kjellmer, G. 1994. *A Dictionary of English collocations,* Clarendon Press, London.

L'Homme, M.C. and Bertrand, C. 2000. "Specialized Lexical Combinations: Should they be Described as Collocations or in Terms of Selectional Restrictions", In *Proceeding of Euralex,* Stuttgart University, pp. 497-506

Martin, W. 1992. "Remarks on Collocations in Sublanguages", In: *Terminologie et traduction 2-3*, pp. 157-164

Mel'čuk, I. A. et al. (1984 and 1988). Dictionnaire explicatif et combinatoire du fransais contemporain: Recherche lexico-semantique (Volume I, 1984; Volume II, 1988). Montréal: Les Presses de l'Université de Montréal.

Mitchell, T. F. 1971. "Linguistics 'Going on': Collocations and Other Lexical Matters Arising in the Syntagmatic Record". In: *Archivum Linguisticum, Vol. II.*

Moon, R. 1998. *Fixed Idioms and Expressions in English.* Clarendon Press, Oxford.

Sinclair, J. 1987. "Collocation - a progress report". In: Steel, R. and Threadgold, T. (eds.) *Language Topics - Essays in honour of Michael Halliday*, pp. 319-333, John Benjamins, Amsterdam

Svenska Akademiens grammatik. 1999. Stockholm: Norstedts ordbok.

Viberg, Å., Lindmark, K., Lindvall, A. & Mellenius, I. 2002. "The Swedish WordNet Project." In: Proceedings of Euralex 2002, Copenhagen University.

Vossen, P. (ed.) 1998. *EuroWordNet : A multilingual database with lexical semantic networks.* Dordrecht: Kluwer Academic.