Human IT
Tidskrift för studier av IT
ur ett humanvetenskapligt perspektiv

Interactive Acquisition of Terminology Describing Job Applicants in Job Advertisements

by Jens Allwood, Maria Cheadle & Anders Green

Abstract

There are several computer based job classification systems for the labour domain and the purpose of which mainly is to enable search and to collect statistics. The systems typically encode information about the job, the workplace, the employer, the salary, necessary education and experience, duration, hours, need for a driving licence etc. But actual job advertisements, for example such that can be found in the Swedish Job Bank, a set of web pages containing information about job vacancies provided by AMS, the Swedish government agency for labour market activities, often contain requirements that are not included in these classification schemes; the applicant should enjoy working with people, be able to pick up new things quickly or be creative. In an effort to extend existing schemes we have developed a taxonomy of properties like these and a computer tool that can be used to analyse text based on the categories in the taxonomy. The tool can also be used to construct future taxonomies. In this work, we will present the taxonomy, the computer tool and some results based on using the tool.

This paper has three purposes; the first is to present a taxonomy of personal properties of job applicants in job advertisements, the second is to present a tool that can be used both to develop such a taxonomy and to analyse text using the taxonomy, and the third is to describe some results of having used the tool and the taxonomy to analyse job advertisements.

Contents

---

1. Background

Our first task was to find suitable categories to classify words describing job applicants. According to Allwood (1989), several kinds of empirical data can be used to support conceptual analysis and construction of taxonomies, e.g. interviews, intuitions or relevant linguistic material. Given that we wanted our categories to cover a large and varied set of job descriptions, we suspected that intuitions and interviews would not alone provide us with sufficient variation. We therefore decided to develop a taxonomy through a direct analysis of a written corpus of job advertisements written by more than one thousand authors from different labour sectors. The following steps were taken:

i. selection of a large set of relevant descriptive terms. The terms were first selected automatically and then further specified on the basis of manual analysis.
ii. construction of a taxonomy for these terms,
iii. construction of a simple conceptual role – relation and property model to be used as a basis for the construction of the tool built to support the analysis.

Figure 1 below illustrates the relationship between the conceptual model, syntactic categories and the semantic categories role, relation and property.

Figure 1: Relations between concept model, syntactic categories, word classes and an example from the corpus.

Concept type		**Syntactic**	**Word class**	**Example**
Role	⇔	Subject	Nouns	du / you
Relation	⇔	Predicate	Verbs	har / have
Property	⇔	Attribute	Adjectives/Nouns	öppet sinne / an open mind

du	har	egen	bil		och	du	är	över	18	år
you	*have*	*own*	*car*		*and*	*you*	*are*	*over*	*18*	*years*
JC	**Verb**	**W1**	**W2**		**W3**	**JC**	**Verb**	**W1**	**W2**	**W3**

Although a more complicated conceptual analysis would be possible, we decided to focus on relations and properties related to the role of being a *job candidate* (JC). It would however have been possible to find properties for other similar roles in the domain, such as the roles of being a *company*, *product* or *an employer*.

In order to carry out step 1, we used information about syntactic relations and categories to find terms. This relationship is depicted in figure 1. For instance, we use the fact that nouns often function as subjects in sentences to establish the location of verbs and adjectives which in turn indicates what terms are used to describe the job applicant.

The linguistic corpus data we used consisted of 12000 job advertisements, published in the Job Bank (Platsbanken) by the National Labour Market Board (AMS) in Sweden. The advertisements in our corpus were available to the job applicants in March of 1998.

(To the top)

2. Automatically collected keyword phrases

Different corpus methods have been used to support the development of terminology resources since the sixties. Two examples of this are concordances and lexical databases which have become powerful tools to support the manual development of lexicons. The fast growing interest in search facilities on the Internet have made machine readable lexicons with thesaurus information important to provide search based on hierarchical classification, see Lycos which provides a broad coverage online search.

The method used in this work is a statistical filter based on the binomial distribution described further in (Manning 1993). By using automatically extracted keyword phrases we can form a keyword context which enables the user to view interesting parts of the corpus. It is constructed using a statistical filter with a hypothesis test based on the binomial distribution. The filter is then applied on a set of patterns from the corpus which are formed by using pairs of n-grams seen in the corpus. These pairs are selected using two types of information as heuristics. First we select a set of words which we believe are cues for job candidates, words such as *you*, *chef*, *person*. Using these words n-gram pairs are collected from the text, for these pairs we then collect histogram statistics. The second thing we use as a heuristic is the information that finite verbs often are found in phrases from our domain that contain keywords describing the employee (e.g. you *are* a skilled chef, you *have* experience of...). Thus pairs collected from the corpus might look like below in Table 1.

Table 1: Different pairs of n-grams as they might appear in a corpus.

JC	P_1 , VERB	# P_1	P_2	#P_1+P_2	Decision
du	*är en skicklig*	21	*kock med*	1	***Reject***
du	*är en*	34	*skicklig kock*	5	***Accept***
du	*är*	590	*en skicklig*	10	***Reject***
person	*vi letar*	89	*efter har*	7	***Accept***

du	*är*	590	*kreativ*	3	***Accept***
du	*är*	590	*kreativ och*	1	***Reject***

The hypothesis P_1+P_2, that the phrases P_1 and P_2 are significant in this corpus, is accepted if the value of the binary distribution from the count of P_1 and the count P_1+P_2 higher than a certain threshold value (currently 0.02 using 12000 advertisements). Otherwise the pair P_1+P_2 is rejected. In table 1, we accept the cases: *vi letar efter har, är kreativ* and *är en skicklig kock*. Accepting the case *vi letar efter har* does not seem to give any information about cue words that describes an employee, but serves as a pattern for writing dedicated rules and is therefore important. The other two cases both signal some property of the employee and are thus good examples of successful use of the method.

(To the top)

3. Constructing the Taxonomy

The construction of a classification system for the keyword phrases was performed in steps. The first step was to determine a set of appropriate features to describe the applicant's personal properties. After studying some job advertisements in general we had some intuitions about properties concerning *education*, *skills*, *abilities*, *insight*, *experience*, *proficiencies* and properties concerning *personality* and *values*. These types of properties then served as potential categories in the process to augment the set of features in order to get a final set that would distinguish the categories from each other.

We divided the keyword phrases into a first coarse grouping consisting of on the one hand educational and formal requirements and on the other hand a group which we refer to as informal properties. The formal properties in the first group are verifiable in one way or another, by grades, documents, references etc and the informal group captures properties that cannot be quantifiable or verifiable, at least not easily. The informal properties typically describe the applicant's personality rather than his or her professional status.

This grouping may seem a bit coarse but it is productive in the way that this seems to form a common ground for how the job advertisements are constructed. For example, if an employer looks for a medical doctor with communicative skills, the applicant will put forward some formal proof of the first property, whereas the second property is of a more subjective nature, which really presupposes a common understanding of what having »communicative skills« actually means.

Table 2: Formal and informal properties of the applicants.

Formal	**Informal**
4-årigt tekniskt gymnasium *4 years technical secondary school*	flexibel *flexible*
minst 18 år gammal *at least 18 years old*	självständig *independent*
dokumenterad samarbetsförmåga	samarbetsförmåga

documented ability to co-operate *ability to co-operate*

This grouping is nevertheless not good enough. The property of being able to co-operate can be said to be formal if an employer should ask for a document to prove that the applicant possesses this ability, but if the employer does not explicitly ask for any proof in the advertisement, we can argue that it is informal. The formal property can then perhaps be defined in terms of the features *verifiability*, *quantifiability* and *consensus*. The absence of those features would then correspond to the informal property.

Nearly all job advertisements state some requirements about *education*, *skills*, *abilities*, *insight*, *experience* or *proficiencies*. Table 3 below lists some examples from the corpus. Our first attempt was to use the mentioned categories above, but we soon saw that there was simply too much overlap. Should an employer require the applicant to have a degree in medicine it seems unlikely that the employer would not wish the applicant to have at least the insights, experiences, skills or proficiencies of a young and inexperienced medical doctor. We have chosen the categories insight and ability, the first one focusing the cognitive capacity of the applicant and the second the focusing on the practical application. Education is in our perspective a process in an educational context which results in new insights and new abilities. Experience is also a process of gaining insight and abilities over a period of time but through seeing and doing things rather than studying. The acquisition of experience is not limited to an educational context.

The terms describing insights are divided into three categories: *insights through education*, *insights through previous work experience* and *experience from life* and terms describing abilities are divided into five: *ability through education*, *ability through previous work experience*, *general ability*, *cognitive ability* and *special ability*. There is some overlap between the categories describing abilities and the categories describing insights.

Table 3: Examples from job advertisements of insight and ability oriented constructions.

Insight and ability oriented constructions
du ska ha högskoleutbildning inr. konstruktion *you should have a university degree in construction*
du är allmänbildad och praktisk *you have general knowledge and are practical*
B – körkort *driving licence*
Erfarenhet av projektledning är meriterande. *Experience of project leading will be considered an additional qualification*
Du har goda insikter i engelska. *You have a good knowledge of English*
Tjänsten förutsätter förmåga att samarbeta/samverka med övrig personal *The situation requires an ability to co-operate with the rest of the staff*
Du ska behärska engelska i tal och skrift. *You must have a good command of both written and spoken*

(To the top)

4. The Taxonomy

Using the methods incorporated in the tool to be presented below, we constructed a taxonomy to describe personal properties of job candidates, as the employers describe them in vacancy advertisements. The development of the taxonomy went hand in hand with the development of the tool.

Using the concept analysis as described in Allwood (1989) we established some different categories for how the job candidate is described in job advertisements. A more extensive list of categories can be found in the appendix.

Insight. This is a very common category and it seems to be possible to divide insight into three main categories, insight you gain from education, experience or life in general. The last category stems from the fact that this often is an explicit requirement appearing in the advertisements, for instance for some jobs within the health sector, drug rehabilitation etc.

Physically and socially identifying properties. This group of categories consists of these categories: age, physical condition, sex, appearances and other properties of a more formal nature, like nationality and where the person lives, for instance.

Properties concerning motivation and action. When we look at the job advertisements it seems important to possess certain properties that concerns the behaviour in terms of actions and motivation of the employee. Terms which fall into this category are *samarbetsvillig* (co-operative), *effektiv* (efficient) and *hungrig* (hungry).

Properties concerning interaction. In some jobs it seems important to be able to interact with people, and also to be aware that the job concerns interacting with people. Terms that falls into this category are *service-minded*, *social* and *dynamic*.

Ability. The categories in this group of categories are very commonly found in the job advertisements. An ability can be obtained either by education or previous work experience, but it can also be some more generic ability that is required from the employee, for instance some cognitive ability or special ability. The ability required by experience of previous work is often stated explicitly (e.g. *you are skilled, an expert of*) whereas the latter abilities often are stated with terms like *logical, analytical, structured, creative* or *plays an instrument*.

Holistic properties. The properties which fall into the category holistic, are the ones which describe the person as a whole (e.g. strong, suitable, competent).

Attitudes. Another set of properties which the employers use to describe the persons they are looking for are the attitudes one should possess in order to be a successful candidate or worker. Attitudes falls into the categories; general attitude (e.g. *nice*, *curious*), attitude towards work (e.g. *ambitious*, *interested*) and emotional attitude (e.g. *harmonic*, *emphatic*).

Moral attitude. In some advertisements certain moral attitudes are required from the employee, for instance *honesty, loyalty, freedom from prejudice*.

Religion, Lifestyle, Ideology. In some advertisements, typically ones where the employer is a church, a union or another organisation certain properties are necessary, for instance *baptised*, *member of a certain church* (kyrkotillhörighet).

(To the top)

5. The Tool

The interactive tool that we have constructed can be used to create rules for carrying out the following three activities in a particular corpus material:

i. support the process of finding terms and,
ii. support the development of classification systems (in our case to analyse job advertisements)
iii. analyse subsets of the corpus

The tool uses a subset of automatically extracted keyword contexts from a large corpus of job advertisements. The keyword contexts are those phrases that occur immediately after a reference to the job candidate often enough to be accepted by a statistical filter. From these keyword contexts rules may be formed that:

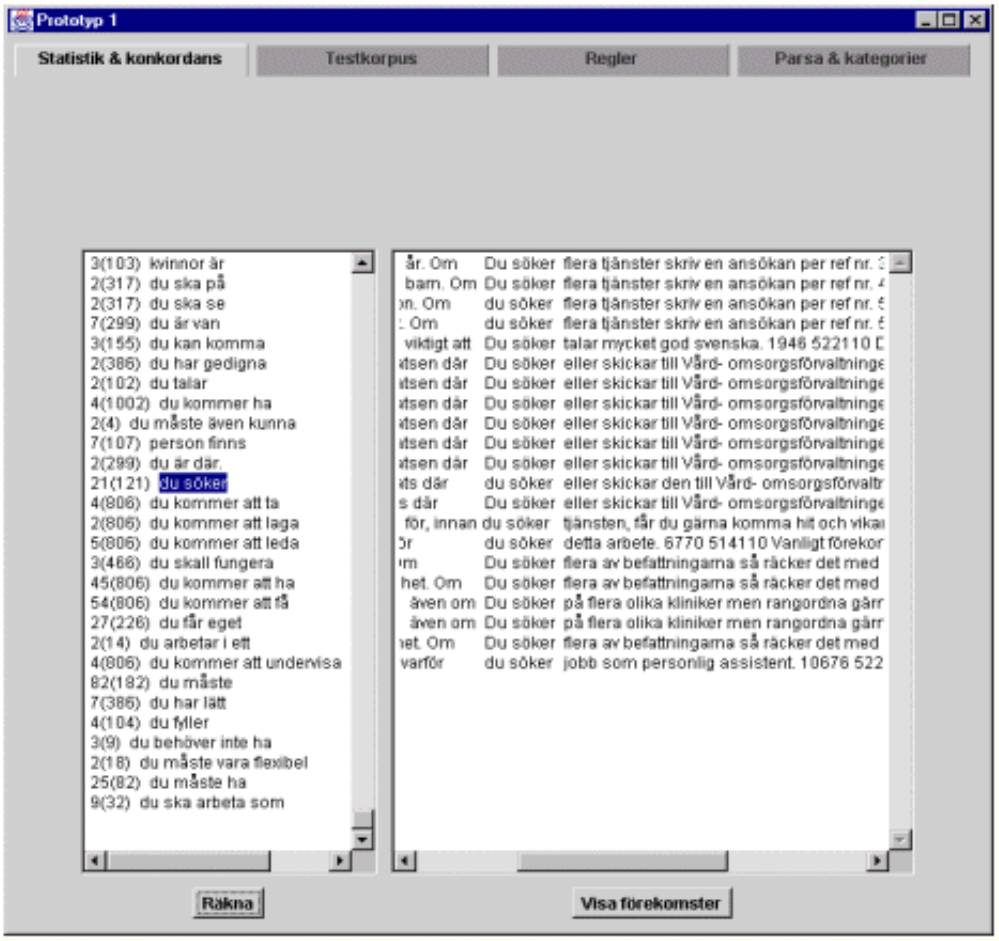
i. assign tags to words,
ii. enable extraction of unknown terms.

The rules are basically context free grammar rules extended to allow for wildcards and named placeholders. A wild card (#) means »any« or »anything«. A named placeholder $X saves every word that appears in the text in the position matching the named placeholder's position in the rule that introduced it. The collected words are associated with the variable used in the named placeholder. The latter is a type of simplistic information extraction.

C1→	you are
C2→	C1 willing
C2→	C1 willing to $C3 and $C4
	The values of $C3 and $C4 are stored in the set of cues for C3 and C4 respectively
C5→	you # # willing to $W
	Covers the case »you would be willing to«, among others, and collects every word matched by $W in W

The rules can then be used in the tool to parse the corpus, as a way of finding other occurences of a particular category in the corpus.

Figure 2: The Statistics and Concordance view



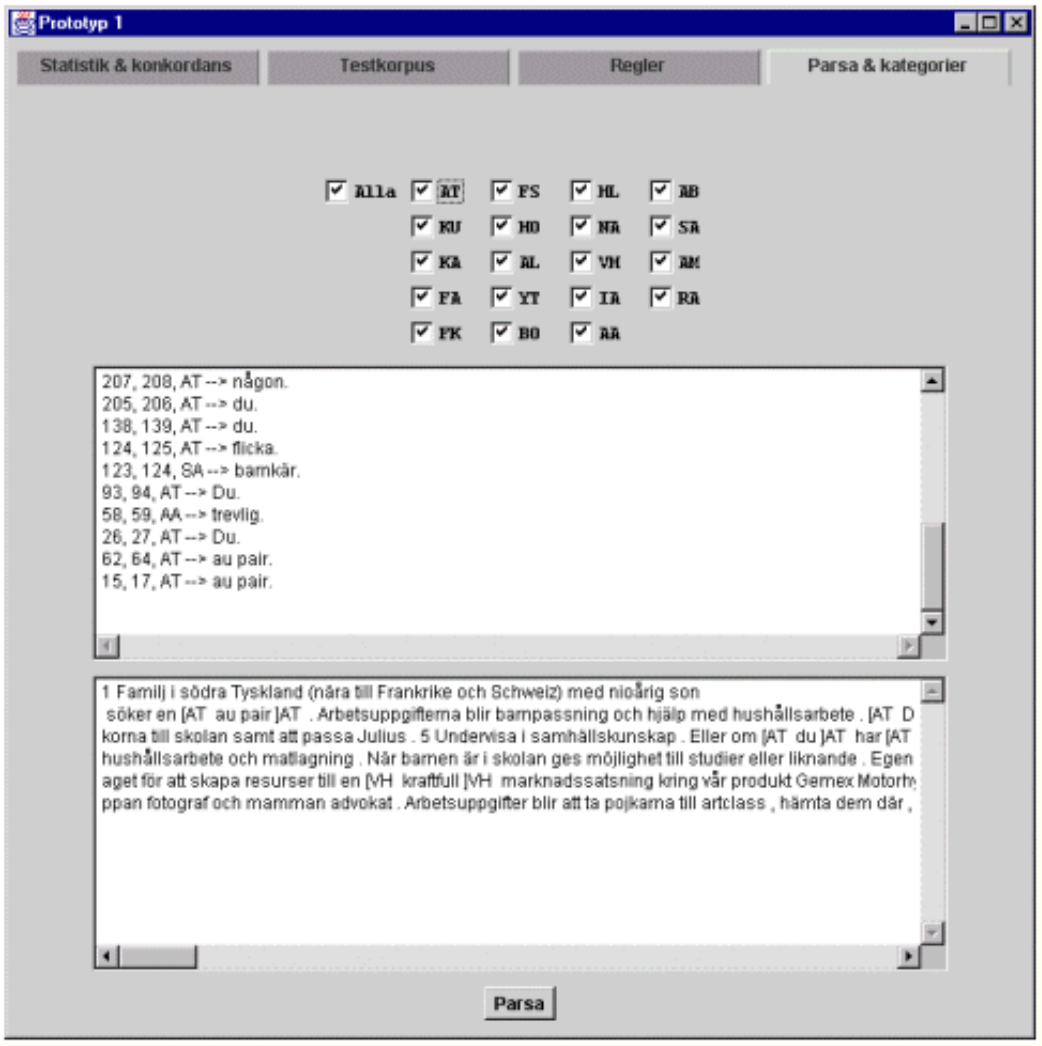
The parser is a bottom up chart parser (cf. Gazdar & Mellish 1989) that has been extended in order to allow the use of wild cards and named placeholders. The tool itself is written in Java, and thus rather platform independent. It has four different views that are used for different tasks:

The Statistics and Concordance view (see figure 2) presents the result from applying the statistical filter algorithm on the corpus. The program uses the 12000 job advertisements, a list of common job candidate words and a list of Swedish verbs in the active form extracted from the Stockholm Umeå Corpus (Ejerhed 1992). The left textarea shows the phrases found in the corpus preceeded by the histogram statistics used as input for the filter. For instance, the row »3 (103) *kvinnor är*« means that the whole token occurs three times out of a total of 103 for the token »*kvinnor*«. By clicking on a row the corresponding concordances will appear in the right text area.

The Corpus view enables the user to choose a corpus to work on and simple editing and file browsing possibilities.

The Rule editing view gives the user the opportunity to specify DCG style rules to analyse the corpus (for an introduction to DCG see Gazdar & Mellish 1989). It offers the same simple file browsing possibilities as the Corpus view.

Figure 3: The Parser and Categories view.



The Parser and Categories view (see figure 3) is used to inspect the result from the rules applied to the test corpus. The user can choose what categories the parses should display by checking or de-checking the category checkboxes which are derived from the grammar and displayed above the result window.

(To the top)

6. Using the tool

From the user point of view there are some remaining issues. Since we know of no tools like this one we have nothing to compare it with in terms of performance. When processing large corpora, today, there are often some points in the interaction when the user has to wait for the program to finish some calculation or parsing process and this is also the case with our tool. This might be partly remedied in a future extension by reimplementing the tool in another programming language. Another solution would be to use a client server architecture, where fast computing would be provided on the server side and the interactive augmention process is done on the client side. However, we have found that the tool works well when the user is iteratively augmenting the rules by analyzing smaller portions of the corpus and adding new rules by using the extraction mechanism. The wildcard feature is also very helpful in the augmenting process.

After using the tool in this fashion we have as a result a collection of rules that capture the expressions used by the authors of the advertisements to describe a certain property.

We have no means to decide when the collection of rules is complete in terms of coverage but as long as we discover new occurences of a category in the taxonomy we know it is incomplete. Below are shown some example parse results after some augmenting of the rules. The initial code in each advertisement is the job classification code from AMSYK.

Table 4: Some parse results.

i. **512320** Vi söker servitörer / servitriser som förutom bordsservering [*AE kan*]*AE* underhålla gästerna med sång och musik . Kanske en språngbräda för [*JC Dig*]*JC* som är på väg in i nöjesbranschen . Vi söker även [*JC dig*]*JC* som inte har [*JC någon*]*JC* musikalisk [*GA [CA förmåga] CA*]*GA* men har [*AW [AW intresse] AW av serveringsjobbet i*]*AW* sig.

ii. **213130** Vi söker mer än en systemutvecklare . Kanske är [*JC du*]*JC* en [*IW erfaren IT*]*IW* - agent som vill flytta fram positionerna . Eller en [*MA målmedveten*]*MA* rookie som vet din [*GA [CA förmåga] CA*]*GA* och potential , men ännu inte fått chansen att visa den? [*JC Du*]*JC* känner [*JC dig*]*JC* hemma i både styrelserum och koja. [*JC Du*]*JC* vet att föra [*JC dig*]*JC* i både smoking och non - smoking . [*JC Du*]*JC* [*AE kan*]*AE* ta snabba och viktiga beslut även i trängda situationer . Här på FÖRETAGET blir [*JC du*]*JC* en av 27 personer på ett mycket expansivt Skellefteåkontor med Sverige som marknad. [*JC Du*]*JC* kommer att jobba med utveckling av produkter och system som format morgondagens IT - lösningar och gör våra kunder konkurrenskraftiga även på en global marknad.

iii. **513210** Vård och omsorg om [*PR boende på kommunens*]*PR* servicehus och gruppboende.

Some immediate reflections on this result are that we should add all the job titles used in the existing AMS search facility to capture as many job candidate references as possible. By using a morphological stemmer we would not need to predict which forms are likely to occur. This would help to keep the number of rules low. If we were to use a part-of-speech tagger on the corpus before analysing it with the tool, we could add a syntactic layer to the collection of rules. Such a layer would certainly help us write more general rules, which in turn would enhance the extraction of new keywords.

(To the top)

7. Using the taxonomy

The collection of rules we compiled manually and later by using the tool were used to analyse the job description part of the corpus. As we have seen, the tool is meant to support the process of constructing classification systems and extracting terminology rather than being a fully-fledged high capacity corpus tool. Therefore, in order to find occurences of the keywords a simple batch program was written, using a reduced version of the rules, with the special purpose to compute statistics on the corpus.

The next step was to compare the statistics from the corpus analysis with the AMSYK classification of advertisements already made in the database by AMS. Table 5 presents the top level of the AMSYK classification of occupational types.

Table 5: The top levels in the AMSYK work classification hierarchy.

Class	Description

0	Armed forces
1	Legislators, senior officials
2	Professionals
3	Technicians and science professionals
4	Clerks
5	Service workers and shop sales workers
6	Skilled agricultural and fishery workers
7	Craft and related trades workers
8	Plant and machine operators and assemblers
9	Elementary occupations

Through this comparison, we have discovered some interesting features. Table 7 in the appendix shows the detailed results of the corpus analysis. Table 6 renders a simplified overview of these results. Table 7 in the appendix shows the detailed results of the corpus analysis. Table 6 renders a simplified overview of these results. (Excluding the armed foreces, C1-C9 correspond to the nine occupational types in table 5.)

Table 6: The simplified results of the corpus analysis. (The categories sum to more than 100%, since there is an overlap between the types of personal property represented in the table.)

	C1	C2	C3	C4	C5
Insights	10.0 %	15.0 %	11.0 %	11.0 %	6.2 %
Abilities	16.0 %	19.0 %	14.0 %	16.0 %	16.0 %
Physically ind. prop.	4.2 %	4.5 %	6.6 %	4.1 %	16.0 %
Motivation/Action	6.3 %	6.7 %	5.9 %	4.3 %	5.8 %
Interaction	2.5 %	0.8 %	0.7 %	1.0 %	1.5 %
Holistic	3.7 %	2.7 %	1.4 %	2.4 %	2.8 %
Attitudes	6.6 %	6.9 %	7.5 %	6.7 %	8.4 %
Candidate	60.0 %	56.0 %	58.0 %	63.0 %	51.0 %
	109.3 %	**111.6 %**	**105.1 %**	**108.5 %**	**107.7 %**

	C6	C7	C8	C9	Overall
Insights	21.0 %	12.0 %	13.0 %	2.7 %	16.0 %
Abilities	3.5 %	19.0 %	18.0 %	10.0 %	10.0 %
Physically ind. prop.	11.0 %	4.1 %	2.6 %	32.0 %	9.9 %
Motivation/Action	2.4 %	6.0 %	5.0 %	4.5 %	6.0 %
Interaction	0.0 %	0.2 %	1.0 %	0.4 %	1.0 %

Holistic	4.7 %	4.7 %	3.0 %	0.5 %	3.0 %
Attitudes	12.0 %	6.6 %	7.9 %	6.8 %	7.4 %
Candidate	55.0 %	56.0 %	61.0 %	45.0 %	56.0 %
	109.6 %	***108.6 %***	***111.5 %***	***101.9 %***	***109.3 %***

Columns C1-C9 show how the different categories of the taxonomy of personal properties of job applicants are distributed within the set of found keywords in each of the nine AMSYK categories. The last column shows the average for the whole corpus. Since the categories of the taxonomy are not mutually exclusive the sums exceed 100%.

Our intuitions, formed when manually scanning the advertisements, that jobs that require more skills are described with longer advertisements, finds some support in the fact that the advertisements in category 1 and 2 are considarebly longer than the average length for the rest of the material. Table 7 in the appendix shows the words per advertisment counts for each category and the ISCO-88 classification of qualifications.

We find it interesting that abilities gained through education are about as infrequent in category 1 and 2 as in category 9 (see Table 7 in the appendix). It is possible that the jobs themselves in these categories carry some implicit information as to what abilities are required, but we have not studied this further.

Another interesting feature is that properties concerning motivation and action are rather evenly distributed over the categories (see Table 6). It would seem that properties of this sort are important to all of the authors.

Expressions referring to properties concerning interaction are most frequent in category 1 (see Table 6). This category covers managers and leaders and we assume that the authors of the advertisements find this a very important category.

References to the job candidate are very common everywhere in the data (see Table 6 in the appendix). 56% of the found keywords were references of this sort. In category 9, however, only 45% of the keywords were references to the job candidate. Our intuitions tell us this is so because the advertisements in category 9 describe jobs that do not require any particular training and tend to focus more on minimal requirements in terms of physically individuating properties. This is supported by the fact that references to the job candidates' place of residence amount to 17% of the keywords in category 9 (see Table 6 and Table 7 in the appendix). The place of residence is more pronounced in the advertisements in category 5,9 and to some extent category 6 (see table 7).

We are aware of some weaknesses in the analysis, these weaknesses stem chiefly from the fact that we used a reduced version of the collection of rules. The collection of rules itself would, if it were possible to analyse the whole corpus with it, most certainly perform better than the simple keyword analysis, especially at disambiguating occurences.

Keywords occur in more than one category which yields a certain amount of overlap. This is not really a problem when the occurences themselves are ambiguous. If we have no method of deciding which category is appropriate we might as well take both. The phrases that contain information about education are most of the time an example of this,

but not always, the same phrases might very well occur in a different meaning, not as a property the applicant should possess.

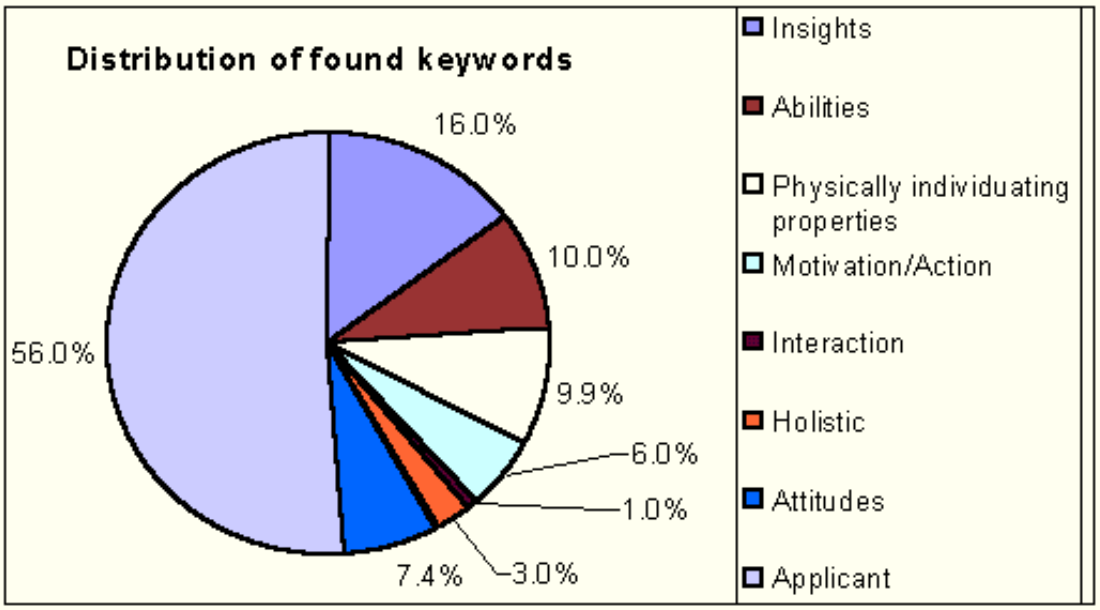
The word *boende* (occupant, housing) in the third of the advertisements (see table 4) is such a word. Since *boende* in this case is not a physically individuating property of the job candidate, this is an error. We can suspect that this word is more frequent in advertisements that offer jobs in a home for the elderly than in other advertisements, an thus we can expect it to occur more frequently in category 5. So when we look for this keyword when we try to find physically individuating properties we will also find the occurences that are not examples of physically individuating properties. This is possibly a part of the explanation why the sum for the physically individuating properties in column number 5, representing category 5 of the AMSYK, hierarchy is considerably larger than that of its neighbours (see Table 6).

(To the top)

8. Concluding remarks

Supposing we wanted to build a system for extraction of information using the taxonomy we have described above, how should we set our priorities? The trade-off between what is needed in order to provide an adequate search facility and what amount of work you would want to spend on developing extraction rules makes it necessary to decide where to put the effort.

Diagram 1: The simplified overall distribution of found keywords.



In diagram 1 we see the simplified overall distribution of found keywords (for a more detailed view see table 7). The most frequently occurring categories are those that concern the insights and abilities the job candidate should possess (e.g. the categories IE, IW, AE, AW). What this diagram does not show is that the occurences of expressions categorized as *Insights*, *Abilities*, *Physically Individuating properties* or *Motivation and Action* can be found in all the AMSYK categories. This is of practical importance when potential users search in the database in terms of recall. Putting effort on rules for the categories mentioned above therefore seems like a good idea, whereas

effort spent on categories like *Interaction* might be less important in terms of search recall.

However, it is important to notice that categories occurring only a few times might convey crucial information of a particular type of job, applicable to no other types of jobs.

For instance, for almost all jobs the religion of the candidate is of no importance, but when a church wants to hire a public relations person creed is important.

It would be interesting to explore the corpus using the same methods as described above but focusing on another role, for instance the employer or the company. Another study which could be made would be to apply this method on descriptions written by the candidates themselves, for instance by looking at CVs to compare the terms used by the employers and the candidates respectively.

Another interesting area to try the tool described in this work would be to explore the possibilities of enable self-organising lexicons by using the rules for lexical extraction. See (Holmes-Higgins & Ahmad 1996).

(To the top)

About the authors

Jens Allwood är professor i lingvistik vid Göteborgs Universitet och dessutom ordförande för Kollegium SSKKII, ett tvärvetenskapligt centrum för kognitionsvetenskap. Hans intressen omfattar bl. a. semantik, pragmatik och studier av kognitiva system.

Maria Cheadle studerar datalingvistik vid Göteborgs Universitet. Hennes pågående examensarbete på SICS (Swedish Institute of Computer Science) handlar om hur vissa språkliga tvetydigheter kan lösas med hjälp av underspecificerade semantiska representationer.

Anders Green är doktorand vid KTH (NADA/IPLab). Hans forskningsområde är människa-datorinteraktion med naturligt språk för styrning av autonoma servicerobotar.

(To the top)

References

Allwood, J. (1989). *Om begrepp, dess bestämning och konstruktion*. Unpublished manuscript.
Arbetsmarknadsstyrelsen (1997). *AMS yrkesklassificering AMSYK och Standard för svensk yrkesklassificering (SSYK).* AMS Förlagsservice
Gazdar, G and Mellish, C (1989). *Natural Language Processing in Prolog*. Addison-Wesley.
Ejerhed, E (1992). *The Linguistic Annotation System of the Stockholm-Umeå Corpus Project*. Umeå University, Department of General Linguistics.
Manning, C. D. (1993). Automatic Acquisition of a Large Subcategorization

Dictionary from Corpora. In *Proceedings of the 31st ACL, Association for Computational Linguistics.*

Appendix

The taxonomy

All the categories with some sample words.

Property	Tag	Sample words
Insikt *Insight*		
Insikt genom utbildning Insight through education	IE	*utbildning inom arbetsområdet*
Insikt genom tidigare arbetserfarenhet Insight from previous work experience	IW	*erfarenhet inom arbetsområdet, erfaren, van, yrkesvana*
Livserfarenhet Experience from life	IL	-
Förmåga *Ability*		
Förmåga genom utbildning Ability through education	AE	*utbildad, kan, utbildning inom, yrkesutbildning*
Förmåga genom tidigare arbetserfarenhet Ability through previous work experience	AW	*yrkesvana, van, praktisk erfarenhet, förmåga*
Allmän förmåga General ability	GA	*snabbhet, kapabel, lätthet, kvalificerat, praktisk*
Kognitiv förmåga Cognitive ability	CA	*logisk, analytisk, systematisk, strukturerad, förståelse*
Speciell förmåga Special ability	SA	*konstnärlig, verbal, händig, händighet, säljarinstinkt, allmänbildad, klurig, idérik, fantasifull, fantasirik, simkunnig, spelar ett instrument*
Fysiskt och socialt identifierande egenskaper *Physically and socially identifying properties*		
Ålder Age	YR	*ung, unga, yngre, 25-30 år, över 45*
Yttre/utseende Appearance/looks	AP	*välvårdad, representativ*
Bostadsort	PR	*bosatt i Helsingborgs kommun*

Place of residence		
Hälsa Health	HL	*fullt frisk, (icke-rökare)*
Kön Sex	SX	*kvinna, man, tjej, dam, kille*
Nationalitet Nationality	NL	*svensk medborgare*
Egenskaper rörande vilja och handling *Properties concerning motivation and action*		
Vilja och handling Motivation and Action	MA	*initiativrik, villig, samarbetsvillig, målmedveten, målinriktad, sugen, hungrig, driftig, drivande, dynamisk, kraftfull, handlingskraftig, resultatinriktad, metodisk, välorganiserad, effektiv, aktiv, hjälpsam, kreativ, produktiv, målorienterad, uthållig, självständig, professionell, pålitlig, ansvarsfull, ansvarstagande, plikttrogen, väluppfostrad*
Egenskaper rörande interaktion *Qualities concerning interaction*		
Interaktion Interaction	IC	*Serviceinriktad, servicemedveten, serviceminded, kundorienterad, säljinriktad, säljande, övertygande, entusiasmerande, social, kamratlig, tydlig, konstruktiv*
Holistiska egenskaper *Holistic properties*		
Holistiska egenskaper Holistic properties	HO	*duktiga, stark, stabil, trivas, lämplig, personlig lämplighet, känslig, framgångsrika, speciell, bra, fantastisk, fysisk, kompetent*
Attityder *Attitudes*		
Allmän attityd General attitude	AG	*trevlig, lättsam, lugn, vänlig, ödmjuk, nyfiken, öppen, lyhörd, spontan, harmonisk, trygg, tålmodig, positiv, pigg, lätt, smidig, envis, viljestark, beredd, flexibel, utåtriktad, förstående, seriös, ordentlig, omdömesgill*
Attityd till arbete Attitude towards work	AW	*ambitiös, noggrann, ordningsam, kvalitetsmedveten, intresserad, motiverad, energisk, entusiastisk, brinnande, engagerad*
Speciell attityd Special attitude	SA	*sportintresserad, sportig, miljöengagerad, kulturintresserad, levnadsglad, barnkär, ungdomlig, mogen, intellektuell, modern, medveten, djurvän*
Känslomässig attityd	EA	*glad, godmodig, gladlynt, glatt,*

Emotional attitude		*kärleksfull, empatisk, omtänksam, rädd, varm, harmonisk*
Moralisk attityd Moral attitude	MA	*ärlig, ärlighet, lojal, hederlig, tolerant, vidsynt, samvetsgrann, fördomsfri*
Religion, Levnadssätt, Ideologi Religion, Lifestyle, Ideology	RL	*vara döpt, konfirmerad i svenska kyrkan, kyrkotillhörighet*

Table 7: Detailed results of the corpus analysis

	CAT1	CAT2	CAT3	CAT4	CAT5	CAT6	CAT7	CAT8	CAT9	*Tot.*
IE	3.5	5.6	4.9	3.2	3.0	1.2	3.8	5.3	0.7	*4.2*
IW	6.9	9.2	6.5	7.7	3.2	2.4	8.4	7.2	2.0	*6.0*
IL										
AE	6.3	10.0	11.0	10.0	14.0	12.0	15.0	13.0	9.5	*11.0*
AW	3.6	3.1	1.3	2.8	0.9	1.2	2.2	3.0	0.4	*1.8*
GA	3.7	2.8	1.2	1.3	0.7	2.4	1.2	1.0	0.1	*1.5*
CA	3.7	2.4	1.0	1.3	0.6	1.2	1.2	1.0	0.2	*1.3*
SA		0.06	0.1	0.2		1.2	0.2	0.3		*0.1*
YR	2.2	3.1	3.4	2.4	3.1	4.7	1.9	1.0	6.8	*4.0*
AP			0.02		0.03					*0.01*
PR	0.7	0.5	2.2		7.4	3.5	0.2	0.3	17	*3.9*
HL			0.05		0.07					*0.03*
SX	1.2	0.8	1.0	1.7	5.0	2.4	1.9	1.3	1.3	*1.9*
NL		0.06	0.02				0.2			*0.03*
MA	6.3	6.7	5.9	4.3	5.8	2.4	6.0	5.0	4.5	6.0
IC	2.5	0.8	0.7	1.0	1.5		0.2	1.0	0.4	*1.0*
HO	3.7	2.7	1.4	2.4	2.8	4.7	4.7	3.0	0.5	*3.0*
AG	3.6	3.1	3.7	4.5	4.2	3.5	2.2	2.3	4.0	*3.7*
AW	2.2	2.9	3.1	1.8	3.1	3.5	2.8	3.0	1.0	*2.7*
SA	0.7	0.8	0.6	0.1	0.3	2.4	1.2	2.0	1.0	*0.6*
EA		0.1	0.07	0.3	0.8	2.4	0.3	0.7	0.8	*0.3*
MA			0.02		0.03					*0.01*
RL										
JC	*60*	*56*	*58*	*63*	*51*	*55*	*56*	*61*	*45*	*56*
KEYWORDS	*808*	*3390*	*4326*	*1002*	*3016*	*85*	*580*	*304*	*1433*	*14944*
KEYWORDS /WORD (%)	*4.0*	*3.3*	*3.5*	*3.2*	*4.0*	*2.3*	*3.2*	*3.1*	*3.6*	*3.5*
ADVERTISEMENTS	*324*	*2308*	*3113*	*870*	*3049*	*164*	*663*	*373*	*1133*	*11997*
WORD/ ADVERTISEMENT	*62.5*	*45.2*	*40.1*	*35.6*	*25.0*	*22.5*	*28.0*	*26.3*	*34.9*	*35.7*
ISCO skill	*4*	*3*	*2*	*2*	*2*	*2*	*2*	*2*	*1*	-
WORDS	*20238*	*104158*	*124731*	*31002*	*76324*	*3686*	*18571*	*9796*	*39558*	*428064*

(To the top)

Return to Human IT 2/1999