Using vacancy mining for validating & supplementing labour market taxonomies

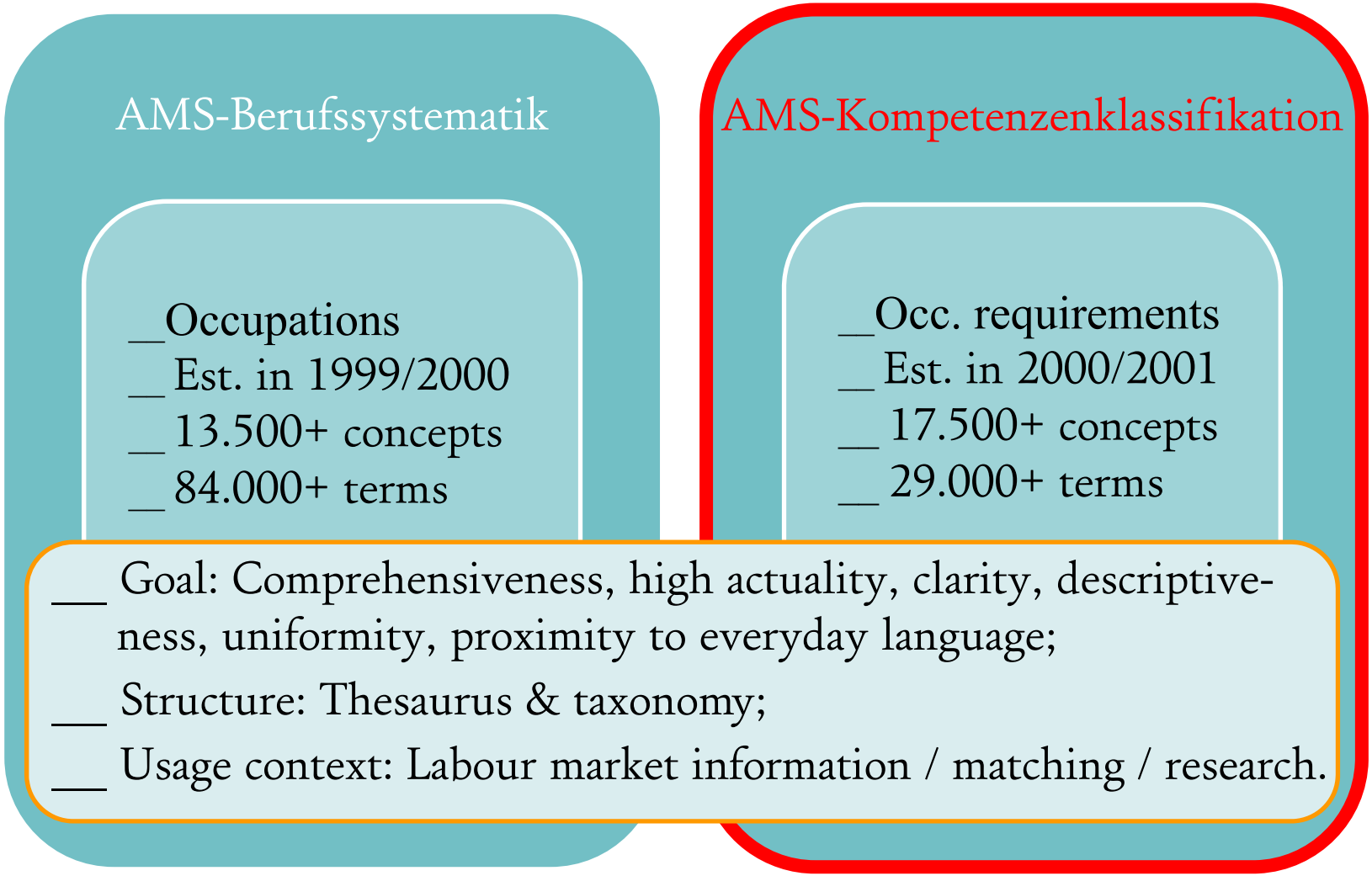
Claudia Plaimauer
3s Unternehmensberatung GmbH
www.3s.co.at

Semantics Conference 2018
Vienna, 12/09/18

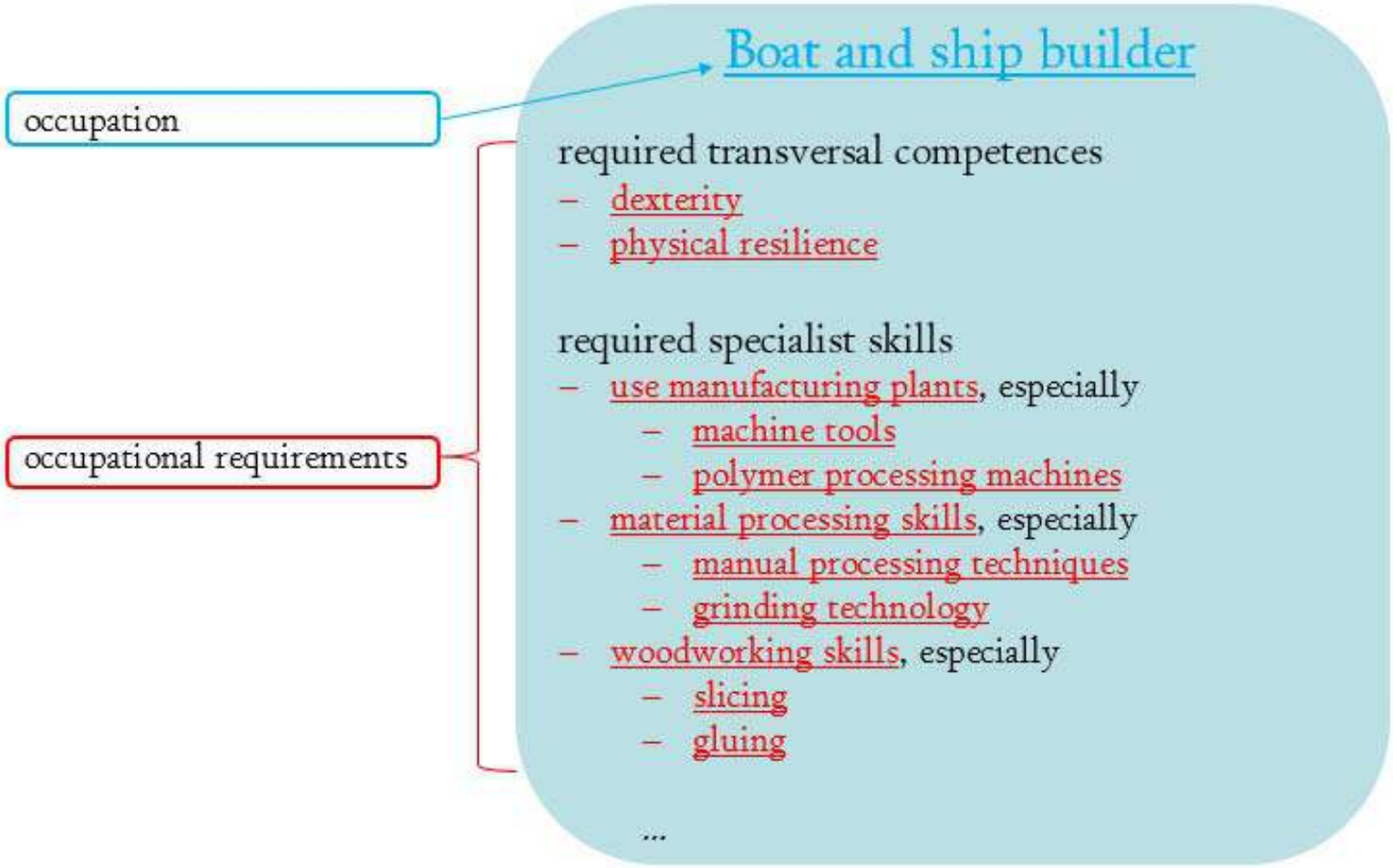
Use of AI and Big Data at 3s

- 2013: 3s & Textkernel (www.textkernel.com) test automatised normalisation of free text survey results (occupations, skills&competences, training needs);
- 2014/15: 3s tests semantic technologies for validating occupational skills profiles (in the context of Cedefop‘s mid-term skills supply and demand forecasts);
- 2015: Jobfeed AT (www.jobfeed.com/at/home.php) goes online (big data platform for querying the Austrian online job market in comprehensive and systematic manner);
- 2017: Pilot project to test potential of semantic technologies for taxonomy maintenance tasks;
- 2017 & 2018: Analysis of Austrian online vacancy market (based on data from Jobfeed); results implemented in AMS Skills Barometer (bis.ams.or.at/qualibarometer).

The Austrian PES' central LM taxonomies



...their interlinkage in occupational profiles



...their maintenance

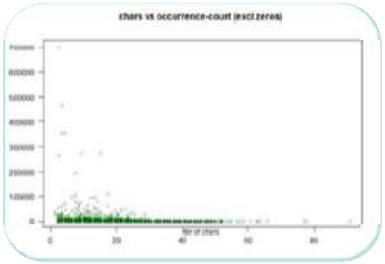
Impulse for amendments comes from

- Expert and non-expert users of these taxonomies;
- Guided, but also spontaneous feedback;
- User-independent quality checks;

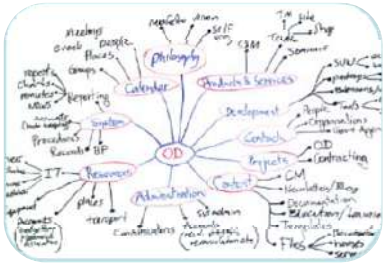
Techniques used in maintenance

- Editorial evaluation of user input/feedback;
- Functional analysis;
- Gap analysis;
- Semantic analysis;
- Terminology control;
- Computer-assisted evaluation of vacancy text. 

Testing AI-based methods for taxonomy management: Goals and expectations



Validation

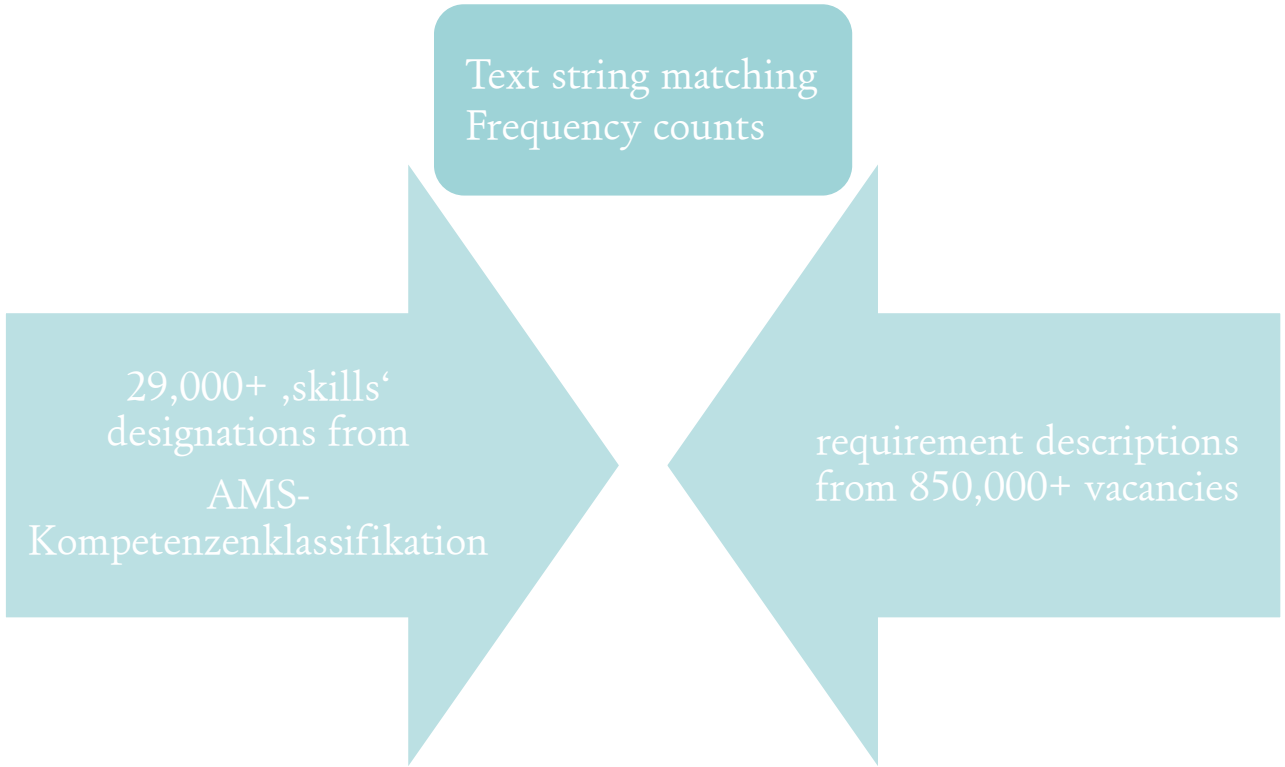


Enrichment

Insights

Savings

Validation of ‘skills’ terms: Method used by Textkernel



Validation of ‘skills’ terms: Results

1. 56% of AMS-Kompetenzenklassifikation‘s ‚skills‘ terms never appeared in vacancies;
2. Negative correlation between term length and frequency of occurrence;
3. Frequency distribution of +/- appearance in vacancies barely differed between preferred and non-preferred terms...
4. ...but some non-preferred terms (NPTs) occurred much more frequently than their affiliated preferred term (PT);
5. Some sub-sections of the ‚skills‘ taxonomy are closer aligned with the language of recruiters than others.

Validation of 'skills' terms: Results - some examples

No occurrence in vacancies, e.g.

- *Maintain and repair construction machines*
- *Determine prices (basic skills)*
- *Impregnate animal pelts against vermin*

Frequency > 60.000:

- *Occupational experience*
- *Knowledge of German*
- *Willingness to travel*

Some non-preferred terms (NPT) occur more frequently than preferred terms (PT):

- NPT *Neukundengewinnung* (F=1.778) - PT *NeukundInnen-akquisition* (F=50)
- NPT *land use planning* (F=239) – PT *Competence in land use planning* (F=0)
- NPT *Sales* (F=54.227) – PT *Sales-related competences* (F=0)

Testing AI-based methods for taxonomy management: Significance of results

Frequency of occurrence in vacancies is only one (and not even the most important one) of several reasons for including a term into the taxonomy, because

- The skills taxonomy does not duplicate but interpret and structure the 'reality' of the labour market; it aims at building a comprehensive model of this specific knowledge domain, and thus also contains elements without any observable labour market relevance within the preceding year.
- It cannot be taken for granted that vacancy text always contains perfectly balanced occupational skills profiles (e.g. concealment of tacitly expected requirements, inflationary use of soft skills).

Testing AI-based methods for taxonomy management: Significance of results - continued

Words/phrases/text segments extracted from vacancy text are frequently

- semantically incomplete once taken out of context, e.g. (*developing?/constructing?/repairing?/using? a) folding press;*
- or too specific to be of general interest, e.g. *support experienced colleague in implementing security measures;*
- or orded in a misleading or vague way, display orthographic and grammatic errors, discriminatory practices, stylistic blunders, etc.,

whereas taxonomy terms aim at broad usability, clarity, descriptiveness and consistency.

→ automatically detected words/phrases/text segments are regarded amendment candidates only, which still have to undergo considerable human processing.

Validation of ‘skills’ terms: Lessons learnt

If ‚AMS-Kompetenzenklassifikation‘ is also used for automated text processing of vacancies, then...

...avoid	e.g.	...promote	e.g.
Excessive pre-coordination of concepts	*Druckmaschinen verkaufen*	Post-coordination	*Druckmaschinen + Verkaufserfahrung*
Explanatory adjuncts in parentheses	*Verkaufspreis ermitteln (Grundkenntnisse)*	Adjuncts in parentheses only for disambiguation of homographs	*ASP (Active Server Pages); ASP (Application Service Providing)*
Excessive contextualisation	*3D-Druck im Automobilbau*	De-contextualisation	*3D-Druck*

Enrichment of vocabulary & conceptual content of the ‚skills' thesaurus: Mix of methods

Automated methods:

- Key word extraction;
- Frequency counts;
- Data cleansing (detection of spelling variants, declensions and typing errors);
- Key word classification;
- Text string matching;
- Co-occurrence analysis.

Editorial methods:

- Identification of additional open source data of related content;
- Exclusion of spelling variants, declensions, typing errors;
- Interpretation of quantitative output;
- Analogous & supplementary searches;
- Semantic analysis;
- Terminology control.

Enrichment of vocabulary & conceptual content of the ‚skills' thesaurus: subsequent editorial processing

- Editorial evaluation of amendment candidates;
- Supplementory enquiry in Jobfeed and other web ressources to clarify content, context and relevance of amendment candidates;
- Terminological adjustment of automatically detected terms to fit prescribed thesaurus format;
- Addition of NPTs, hidden search words, definitions and scope notes;
- Integration of new terms/concepts into semantic structure of taxonomy.

Enrichment of vocabulary & conceptual content of the ‚skills' thesaurus: from automatic detection to editorial integration

Output of 'skills'mining:

- 1.900+ potentially 'new' occupational requirements;
- approx. 900 of these resembled specialist 'skills';
- all 'skills' terms listed with frequency of occurrence, context (most frequently co-occurring occupation) and suggestion for allocation (proposed position in taxonomy).

Result of subsequent editorial processing (focus on specialist 'skills'):

- Addition of 635 terms to the 'skills' thesaurus, of these
    - 366 NPTs;
    - 172 hidden search terms;
    - 97 PTs (= new concepts).

Assessment & Outlook

Text mining is a highly effective method for identifying evidence-based amendment needs for thesauri, but it comes at a price.

→ repeat text mining only at larger intervals.

There is a hard to reconcile tension between controlled vocabulary and natural language, especially

- pre-coordinated (e.g. *Druckmaschinen verkaufen)*
- formally disambiguated terms (e.g. *ASP* (*Active Server Pages*))

hamper the applicability of the thesaurus in automated vacancy coding.

→ Taxonomy should also include formats predominately found in vacancy text as NPTs or hidden search words to improve automated normalisation of requirement text.

Thank you for your attention!

Claudia Plaimauer
3s Unternehmensberatung
Wiedner Hauptstraße 18
1040 Vienna, Austria
Tel +43-1-5850915/33
plaimauer@3s.co.at