Word formation in electronic language resources: state of the art analysis and requirements for the future

Janina Radtke & Ulrich Heid

Keywords: *word formation*, *morphology*, *electronic dictionaries*, *user needs*.

Abstract

We report on a state of the art survey on electronic lexical resources for word formation; these include online specialized dictionaries for interactive use, a grammar information system, as well as a few online tools for morphological analysis. Our comparison is inspired by the Function Theory of Lexicography (e.g. Tarp 2008), and by a definition of needs of users in different communicative situations. Our survey is part of plans towards electronic dictionaries for word formation, and we thus formulate requirements that such dictionaries should ideally fulfil.

1. Introduction

In this paper, we analyse existing online resources for word formation with respect to the data categories they contain, the access to these data categories offered to the user, as well as with respect to general usability aspects. Alongside this state of the art, we derive requirements for future work on the design of online dictionaries for word formation.

Dealing with word formation means to address derivation and compounding, as well as other word formation processes, in terms of word formation elements (affixes, stems etc.), word formation processes and word formation products (derived words and compounds). In this article, emphasis is on derivation; we analyse online resources for different languages[1], as the methodology for the lexicographic presentation of word formation is generalisable across languages.

By online resources, we mean all systems which provide interactively retrievable information about word formation, i.e. online dictionaries, grammatical information systems, and automatic morphological analysers (see section 2.2). We will compare the description and presentation of derivational affixes and their properties, as well as the analysis or generation of complex words and the ways in which users can access information about both (section 3). We also address briefly the usability of the tested systems (section 4) and draw conclusions for future lexicographic work in this domain (section 5).

As far as user orientation is concerned, our work is inspired by the Function Theory of Lexicography (cf. e.g. Tarp 2008)

2. Online information about word formation: needs and available resources

Word formation involves (free and bound) morphemes, i.e. words, their stems, and affixes. Both, derivation and compounding follow rules which have morphological aspects (combinability of elements) and semantic ones (interpretation of transparent complex words).

2.1. *User needs*

User needs with respect to word formation depend on the situation in which a user needs lexicographic data; for cognitive purposes, i.e. to learn about the word formation elements and

rules of a (foreign) language, dictionary grammars or grammar information systems are particularly useful (cf. Tarp 2008: 222ff). In text reception, a major issue is the analysis of complex words: even though many of them are productively built and semantically "transparent", there are many lexicalised exceptions; both the rule-based default interpretation and the idiosyncrasies need to be described. In text production, complex words allow the users to express themselves compactly in a stylistically varied way; they need to be able to ascertain the existence (and use) of complex words, as well as to use word formation processes in order to find appropriate lexical items for the text they are about to produce (cf. Tarp 2008: 221-226).

To satisfy the above needs, a lexical resource must provide a detailed description of affixes and their linguistic properties, base morphemes and their linguistic properties, word formation processes, and complex words. Access to such data should be possible from (almost) any of these elements.

For reception purposes ("R" in table 1), a morphological (and semantic) analysis of complex words and data about base and affix, about the word formation process and the properties of the complex word are needed; production "P", on the other hand, starts from a base and requires data about possible complex words, and the degree of their transparency[2]. Cognitive needs "C" typically include language learning and may be receptively oriented (from affixes to their readings and properties), or productively (the same type of knowledge, plus knowledge about word formation processes). We summarize the above requirements in table 1. We will make reference to the numbered use situations later in this paper.

Table 1. Usage situations (Receptive, Productive, Cognitive) vs. search input and expected output from electronic resources.

#	Situation	Search item(s)	Output
1	R	Word formation product (=WFP)	Properties of WFP, Base, Affix(es), Process
2	R/C	Affix	Properties of the Affix
3	P	Base	WFPs and their properties
4	P/C	Affix	Properties of the Affix, Examples of WFPs
5	P/C	Process	Examples of WFPs

2.2. *Systems under analysis*

The above usage situations are served by different types of online language information tools: (1), (2) and (3) of table 1 are typically dealt with in electronic learner's dictionaries (cf. For example ELDIT[3]), (2), (4) and (5) are best covered by grammatical information systems or dictionary grammars, such as the *grammis* dictionary[4] and the online word formation description offered by canoo.net (based on Word Manager[5]). Many printed dictionaries (such as *LDOCE* or *Cobuild*) contain outer texts devoted to the principles of (2) and (4). Specialized electronic word formation dictionaries, such as the DSVC (*Diccionari de Sufixos Verbalitzadors del Català*, Bernal 2000) and MuLeXFoR (Cartoni/Lefer 2010), mainly cover (1), (2) and (5). Finally, automatic morphological analysers are intended for (1), and

automatic generators or lists of word formation products for (3). Details of the automatic morphological tools we analysed are summarized in table 2.

Table 2. Morphology systems analysed.

Language	Situation	Name	References
DE	1	GerTWOL	Lingsoft (n.d.), Koskenniemi/Haapalainen (1996)
	1	SMOR/ Morphisto	IDS (2008)
	1, 3	WordManager	Canoo Technology (2002), Domenig (1988)
ES	1, 3	GEDLC	GEDLC (1986), Santana (2003)
FR	1	DériF	Namer (n.d.)

3. The treatment of word formation in online resources

In the following analysis, we follow the subdivision of usage situations summarized in Table 1, comparing examples from all types of resources that provide appropriate information.

3.1. *Analysis of word formation products*

In a reception dictionary perspective, obviously complex words need to be part of the macrostructure, and they must be described in the same way and with the same detail (grammatically, semantically, pragmatically) as simplex words. In addition, references (e.g. links) to their bases, to affixes and to word formation processes would help users to identify their morphological status and properties. As one cannot expect users to have full knowledge of word formation processes, Tarp (2008: 223) suggests to cross-link derived words both with their bases and with possible further derived words (e.g. *exclusion* linked to both $exclude_V$ and $exclusionary_{Adj}$) which would allow e.g. language learners to constitute for themselves a full network of morphologically related words.

We are not aware of such systematic links in any online learner's dictionary; ELDIT has general summary tables only. DSVC and MuLeXFoR cover each a well circumscribed, but comparatively small part of word formation, within which they give access, from a complex word, to the process and to the components it is built from. The analysis of complex words is one of the application domains of automatic morphology systems: all of those listed in table 2 provide a decomposition into morphemes. For a large number of complex words, canoo's *WordManager* also provides a description of word structure which has been manually checked by the authors and stored as such. For rare and hypothetical words, the output is produced according to (productive) rules and marked as unverified. DériF in addition provides a decomposition of a complex word into morphemes, but also a paraphrase that allows the user to better understand the underlying word formation process, while GEDLC links to abstract semantic categories. Examples of the analysis output provided by *DériF* and *GEDLC*, for derived verbs (*appauvrir* and *empobrecer*, respectively), are given in figure 1, below:

DériF

appauvrir/VERBE==> [a [pauvre ADJ] VERBE] (appauvrir/VERBE, pauvre/ADJ) " Rendre pauvre"

GEDLC

empobrecer: De significación inmaterial⏩Actos de la voluntad⏩⏩Alteración, mudanza
⏩De estado o condición⏩⏩Alteración, cambio, mudanza

An ideal reception dictionary would provide an analysis of complex words, possibly with their structure or with a paraphrase (as given by DériF) and detailed (standard) lexicographical data. The *WordManager* principle, namely to cover frequent items by means of manually verified analyses and to capture productive word formation by means of rules seems ideal to us. It ensures high output quality and broad coverage at the same time, and it informs the user about the "source" of its analyses.

3.2. *Describing affixes*

Information about affixes and their properties is present in many dictionaries and in all automatic morphology systems.

Relevant properties concern (i) the affix itself and (ii) its selection with respect to bases. The first subset includes the category of the affix, its prosodic properties, possible variants and "competing" affixes, but also its origin (in terms of native vs. neoclassical) and its productivity (in the sense e.g. of Baayen 2000). A description of the selection preferences of an affix should include all those properties of bases on which affixes are selective; examples from German are the category of the base, its origin (native or neoclassical), its morphological form (simplex, complex, abbreviated, ...) and the stem type (in Fuhrhop's (1998) sense: e.g. derivation vs. compounding stem).

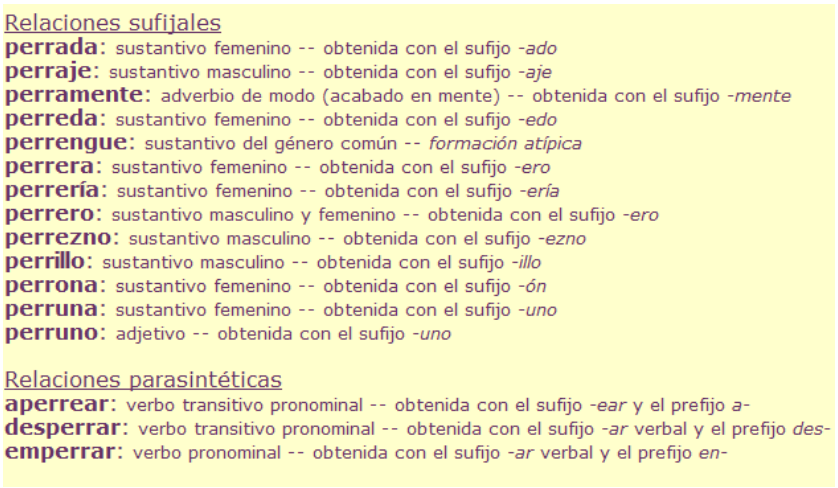
Such a detailed description is only found in the (mainly human-readable) dictionary of the grammar information system *grammis*, and in *canoo's* interactive web pages on word formation. The specialized word formation dictionaries *DSVC* and *MuLeXFoR* give part of the data. The printed dictionary by Gabriele Stein (2007) is also very explicit on many of the above properties, especially also on diasystematic marks of affixes and combining forms.

An ideal, detailed description of the properties of affixes should be accompanied by examples and by links to all (relevant) complex words in the macrostructure as well as to all word formation rules that involve the affix in question. Obviously, the above mentioned broad description of affixes is mostly relevant for cognitive purposes, such as learning word formation principles of a language. It is however also helpful for advanced users to understand whether they can build and use a complex word with a given affix (production) or which exact meaning and connotation a derived word may have (reception).

3.3. *From bases to word formation products*

For text production, Tarp (2008) suggests cross-links between all lexical items related by word formation relations (see above, 3.1). This is obviously only possible for a closed lemma inventory, and thus contrary to word formation productivity. Nevertheless, this proposal can

be followed for frequent items, as is the case in *canoo*, where all complex words related to a given item (and stored in the system dictionary) can be listed. The learner's dictionary ELDIT also provides such lists. Similarly, the Spanish morphology system developed at University of Las Palmas (GEDLC) lists related items, described in terms of their morphemes; some of them are however extremely infrequent (such as *perramente)* or even the product of overgeneration.


Relaciones sufijales
perrada: sustantivo femenino -- obtenida con el sufijo *-ado*
perraje: sustantivo masculino -- obtenida con el sufijo *-aje*
perramente: adverbio de modo (acabado en mente) -- obtenida con el sufijo *-mente*
perreda: sustantivo femenino -- obtenida con el sufijo *-edo*
perrengue: sustantivo del género común -- *formación atípica*
perrera: sustantivo femenino -- obtenida con el sufijo *-ero*
perrería: sustantivo femenino -- obtenida con el sufijo *-ería*
perrero: sustantivo masculino y femenino -- obtenida con el sufijo *-ero*
perrezno: sustantivo masculino -- obtenida con el sufijo *-ezno*
perrillo: sustantivo masculino -- obtenida con el sufijo *-illo*
perrona: sustantivo femenino -- obtenida con el sufijo *-ón*
perruna: sustantivo femenino -- obtenida con el sufijo *-uno*
perruno: adjetivo -- obtenida con el sufijo *-uno*

Relaciones parasintéticas
aperrear: verbo transitivo pronominal -- obtenida con el sufijo *-ear* y el prefijo *a-*
desperrar: verbo transitivo pronominal -- obtenida con el sufijo *-ar* verbal y el prefijo *des-*
emperrar: verbo pronominal -- obtenida con el sufijo *-ar* verbal y el prefijo *en-*

Figure 2. Screenshot from GEDLC showing morphological relations for the word *perro*.

The example of GEDLC shows that a mere listing of word formation products, possible with word formation information, is, alone, not very useful. A more detailed linguistic description is needed in addition. What is not really fully available, but could ideally be provided the same way as the above mentioned lists of complex words (cf. also 3.4), is an interface that allows the user to enter a base, to select a meaning that can be expressed by means of a word formation process and to find out about the complex words available to express that meaning in the language under study. Such a device would exactly fit the needs of text production; its relevance is shown by user logs from searches in dictionaries: Bergenholtz/Johnsen (2005:133) found, for example, morphologically regularised cases such as DK *bekraefigelse* (instead of *bekraeftelse* "confirmation") and non-lexicalised (but plausible) derived words (such as *forspørgelse* instead of *forspørgsel* "request") among the top-500 not-found search words in their online dictionary.

3.4. *From affixes or word formation processes to complex words*

A text production dictionary should relate paraphrases of word formation processes (e.g. "action of V-ing", "result of V-ing") or, alternatively, of affixes (and their readings) with examples of word formation products built by the process, and with the properties of such word formation products. This is done very explicitly (both from a given affix and, fully onomasiologically, from a meaning paraphrase) in DSVC and in the *grammis* system. MuLeXFoR provides in addition such data in parallel for three languages: English, French and Italian.

3.5. *The existing tools: intermediate summary*

So far, our analysis has shown that the specialized electronic dictionaries DSVC and MuLeXFoR, as well as the grammar information system grammis provide the most detailed description of word formation in an electronic resource. They use, however, different terminology: what is called RCM in DSVC (morphological construction rule), is termed LFR (lexeme formation rule) in MuLeXFoR, and just given, without any label, in grammis. Interestingly, all three resources provide roughly the same linking, i.e. the same access path to morphological data: from affixes, users can follow a link to exemplary word formation products; from complex words (i.e. WFPS), there are links to (i) their meaning (paraphrase), to (ii) the affix involved, and (iii) to a paraphrase of or a label for the word formation process in question. From such paraphrases or labels, users are referred to either examples of word formation products build within the process, or to the affixes involved.

Figure 3 shows these access paths (via selection steps, marked as arrows) and the different terminology used in MuLeXFoR (no boxes), DSVC (light grey) and *grammis* (dark grey).

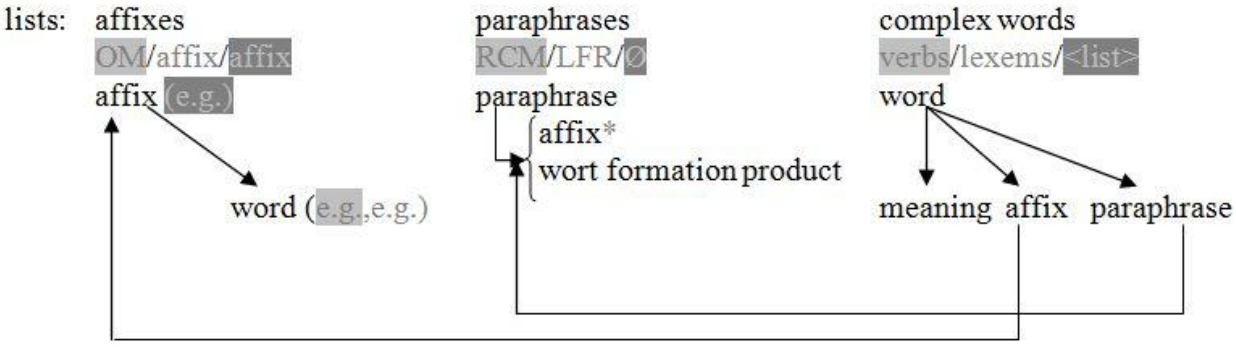


Figure 3. Link structure of DSVC, MuLeXFoR and grammis.

The above short summary of the state of the art shows a number of desiderata:

- there is no general word formation dictionary for text production, yet;
- there is no combination of an interactive dictionary and a morphological analyser[6] for text reception;
- specialised interactive morphological dictionaries are upcoming, but still don't cover a broad range of phenomena (DSVC, MuLeXFoR) and seem to be mostly oriented towards a public of linguists rather than lay persons.

4. Graphical user interfaces for word formation dictionaries

Ideally, Graphical User Interfaces (=GUIs) for word formation dictionaries should optimally support users in the usage situations shown in table 1 or in a relevant subset thereof. Here, we cannot give a full assessment of the existing GUIs; however, a few general remarks are in place:

- some online morphological analysis systems like *GerTWOL* come with no (or only with a very simple) GUI; such systems would need to be integrated into custom-made GUIs for lay users;

- the GUI of DSVC is designed for experts (linguists); already its first page offers three types of search, a list of morphological processes and examples thereof in one screen; with this richness in detail, the dictionary is however hard to use unless one first has a detailed introduction;
- the GUI of MuLeXFoR follows mostly the logic of DSVC, (see fig. 3). It offers two GUI designs, one for experts and one for lay persons; the only difference is terminology: e.g. for usage situation 5, expert linguists search by "lexeme formation rule" and laymen by "meaning";
- despite its generally very user-friendly layout, *canoo* has problems to visualise large numbers of complex words in a single graph;
- the grammar information system *grammis* has flat, textual descriptions of affixes, but little possibility to adapt the lexicographic data it provides to specific user needs.

These few examples may suffice to show that there is a real need for a design of a word formation dictionary that is inspired by user needs and usage situations.

5. Conclusions

In this article, we sketched the state of the art in the design and realisation of word formation dictionaries and/or resources with a substantial component dealing with word formation.
We analysed specialised online dictionaries, grammar information systems and automatic morphological analysers. Interestingly, there don't seem to be integrated products yet (e.g. dictionaries with more sophisticated "grammar" explanations or with automatic analysis tools): even in portals, different functions at most coexist, without interaction. Tools and techniques would from Natural Language Processing however allow us to design integrated tools and to thereby to better serve users.

We also noticed that both coverage and degree of detail vary considerably, and that there is only implicit reference to corpus data. As a consequence, the long-known dilemma between static lexicographic data on the one hand and word formation as a productive, open-ended process that is by nature not easy to capture in a static data collection, remains unsolved, as yet. This article has summarized needs and sketched possible solutions: the implementation remains to be done.

Notes

[1] French, German, Catalan, Spanish, Italian.
[2] E.g. to avoid that learners use a lexicalised derived word in a situation where they would have wanted to use a productively built one: DE *zahlbar* and EN *payable* are not exactly "which one is able to pay", while DE *bezahlbar* has a.o. roughly this meaning ("affordable").
[3] EURAC 1999/ http://dev.eurac.edu:8081/MakeEldit1/Eldit.html
[4] IDS 2000 / http://hypermedia.ids-mannheim.de/grammis/
[5] Domenig (1988)
[6] *canoo* goes some steps into this direction, but more by co-habitation than by integration.

References

A. Dictionaries

***Collins COBUILD English Dictionary for Advanced Learners* 2000.** (3rd ed.) London: HarperCollins. (COBUILD3)

Mayor, M. (ed.) 2009. *Longman Dictionary of Contemporary English*. (5th ed.) Harlow: Pearson Education. (LDOCE5)

B. Other literature

Baayen, R. H. 2001. *Word Frequency Distributions.* Dordrecht: Kluwer Academic Publishers. [=Text, Speech and Language Technology 18]

Bergenholtz, H. and M. Johnsen 2005. 'Log Files as a Tool for Improving Internet Dictionaries.' *Hermes Journal of Linguistics* 34: 117–141.

Bernal, E. 2000. *Els sufixos verbalitzadors del català. Relacions semàntiques i diccionari*. Ph.D. Diss., [en línia] Barcelona: Consorci de Biblioteques Universitàries de Catalunya (CBUC) - Centre de Supercomputació de Catalunya (CESCA).

Cartoni, B. and Lefer, M.-A. 2010a. 'The MuLeXFoR Database: Representing Word-Formation Processes in a Multilingual Lexicographic Environment.' In N. Calzolari, K. Choukri, B. Maegaard, J. Mariani, J. Odijk, S. Piperidis, M. Rosner and D. Tapias (eds.), *Proceedings of the 7th International Conference on Language Resources and Evaluation*. Valletta: European Language Resources Association.

Cartoni, B. and Lefer, M.-A. 2010b. 'Improving the representation of word-formation in multilingual lexicographic tools: the MuLeXFoR database.' In A. Dykstra and T. Schoonheim (eds.), *Proceedings of the XIV Euralex International Congress, Leeuwarden, 6-10 July 2010*. Ljouwert: Fryske Akademy / Afuk.

Domenig, M. 1988. 'Word Manager: A System for the Definition, Access and Maintenance of Lexical Databases.' In D. Vargha (ed.): *Coling Budapest: Proceedings of the 12th International Conference on Computational Linguistics.* Budapest: John von Neumann Society for Computing Sciences, 154–159.

Fuhrhop, N. 1998. *Grenzfälle morphologischer Einheiten*. Tübingen: Stauffenburg. [=Studien zur deutschen Grammatik 57]

Koskenniemi, K. and M. Haapalainen 1996. 'GERTWOL – Lingsoft Oy.' In R. Hausser (ed.), *Linguistische Verifikation*. Tübingen: Max Niemeyer Verlag, 121–140. [= Sprache und Information, 34]

Santana, O., J. Pérez, Z. Hernández, F. Carreras, G. Rodríguez, L. Losada and J. Duque 2003. 'Morfología del español: Reconocimiento y generación automáticos. Desarrollos del Grupo de Estructuras de Datos y Lingüística Computacional de la Universidad de Las Palmas de Gran Canaria (GEDLC).' In Santana et al. (eds.), *Estudios sobre el español de Canarias. Actas del I Congreso Internacional sobre el español de Canarias.*
http://www2.dis.ulpgc.es/~fcarrera/art/art12.pdf

Stein, G. 2007. *A Dictionary of English Affixes. Their function and meaning*. München: Lincom.

Tarp, S. 2008. *Lexicography in the Borderland between Knowledge and Non-Knowledge: General Lexicographical Theory with Particular Focus on Learner's Lexicography*. Tübingen: Niemeyer. [= Lexicographica Series Maior 134]

C. Internet Sites

Bernal, E. 2011. *Diccionari de Sufixos Verbalitzadors del Català*.

http://www.elisendabernal.com/
Canoo Technology 2002. *Wörterbuch der deutschen Wortbildung. Grafische Darstellungen der Wortbildungszusammenhänge für über 200.000 deutsche Schlagwörter und Redewendungen, mit Links zu den entsprechenden Wortbildungsregeln.* http://www.canoo.net/services/WordformationDictionary/ueberblick/index.html
Cartoni, B. and M.-A. Lefer 2010. *MuLeXFoR*. https://sites.google.com/site/mulexfor/
Departamento de Informática y Sistemas de la Universidad de Las Palmas de Gran Canaria 1986. *GEDLC "Grupo de Estructuras de Datos y Lingüística Computacional."* http://gedlc.ulpgc.es/investigacion/scogeme02/lematiza.htm
EURAC 1999. *ELDIT "Elektronisches Lernerwörterbuch Deutsch Italienisch"* http://dev.eurac.edu:8081/MakeEldit1/Eldit.html
IDS 2000. *grammis - Das grammatische Informationssystem des IDS*. http://www.ids-mannheim.de/grammis/
IDS 2008. *Morphisto – Der Lematisierer des IDS für TextGrid*. http://ingrid.sub.uni-goettingen.de/cgi-bin/analyze.cgi
Lingsoft: GerTWOL (n.d.). *Morphologisches Analysesystem für das Deutsche*. http://www2.lingsoft.fi/cgi-bin/gertwol
Namer, F. (n.d.). *MorTAL DériF – Dérivation en Français*. http://www.cnrtl.fr/outils/DeriF/