Towards a lexical database of Dutch taboo language

Gerhard B van Huyssteen[1], Carole Tiberius[2]

[1] Centre for Text Technology (CtexT), North-West University, Potchefstroom, South Africa
[2] Instituut voor de Nederlandse Taal, Leiden, The Netherlands
E-mail: Gerhard.VanHuyssteen@nwu.ac.za, Carole.Tiberius@ivdnt.org

Abstract

Over the past 45 years, at least eighteen Dutch paper-based dictionaries of taboo-language (or taboo-related language) have been published (i.e., as visible works of lexicography). However, none of these are available as (linked) lexical data that could be integrated in natural language processing (NLP) tools and applications (i.e., as invisible works of lexicography). In this paper, we describe the development of a comprehensive lexical database of taboo language (LDTL) for Dutch (TaboeLex) that can be integrated in NLP tools and applications. TaboeLex will be made available as open data, i.e., as a freely available, structured, annotated lexicon that can be linked to other data in the future. The paper focusses on the first phase of the project, namely, to define and design TaboeLex.

Keywords: Dutch; lexical database; swearword; taboo language

Warning: This paper contains content that may be offensive or upsetting.

1. Introduction

Despite giant strides that have been made over the past thirty years in digitalising and automating lexicographic work, resources for specialised purposes and non-mainstream languages are still often neglected. As a case in point, even though at least eighteen Dutch paper-based dictionaries of taboo words (see 2.1 for a definition) have been published over the past 45 years (i.e., as visible works of lexicography), none of these are available as (linked) lexical data that could be integrated in natural language processing (NLP) tools and applications (i.e., as invisible works of lexicography).

Lexical databases of taboo language (LDTLs) are specialised digital resources that could be used as sources of linguistic and extralinguistic knowledge in many natural language processing (NLP) systems (see 2.2). Although such an LDTL could be simply a wordlist, for our purposes we consider an LDTL a digital collection of linguistic constructions that has been annotated or enriched in some way (e.g., with part-of-speech information, offensiveness ratings, meanings), and that is structured (e.g., encoded in XML). Most often, the primary use of LDTLs is to recognise words that could be potentially offensive to a specified community of language users (e.g., children). Despite their immediate practical value, and despite the fact that "much work has been done on abusive language detection in general", much remains to be learned about "lexical knowledge for the detection of abusive language" (Wiegand et al., 2018), as

well as about the development and implementation of LDTLs for languages other than English.

In this paper we will report on the first phase of a project[1] to develop a Dutch LDTL (**TaboeLex**) consisting of potentially offensive constructions (words, word groups, expressions) as linked open data (i.e., a freely available, structured, annotated lexicon that could be linked to other data in future). In section 2, we will give a definition of what we mean by taboo language, and we will set the scope of TaboeLex. Section 3 then describes the design of the database. Section 4 concludes the paper, outlining future work.

2. Definition and scope of TaboeLex

2.1 Taboo language

Referring to the term *swearing*, Stapleton et al. (2022: 2) point out that "precise definitions and criteria are sometimes difficult to pin down [..., e.g.,] whether swear words can be used with literal (as opposed to figurative) meaning". For purposes of this project, we define *taboo language* as linguistic constructions that are potentially offensive to some users in some contexts; constructions are form-meaning pairings on a morphological, lexical or syntactic level (see Goldberg (2006) for an extended view). We therefore use *taboo language* as a hypernym to include other phenomena and/or synonyms like *swearing, cursing/cussing, maledicta, profanity, blasphemy, obscenity, vulgarity, euphemisms and dysphemisms, verbal abuse, verbal sparring*, (*racial*) *slurs*, *terms of abuse, insults, offensive language, dirty language*, etc.

Our definitions and categories are all based on an extensive review of literature from various disciplines that aim to define taboo language, identify types of taboo language, sources of taboo language, etc. Most influential were Hirsch (1985), Hoeksema (2019), Jay (2018), Jay and Janschewitz (2008), Lewandowska-Tomaszczyk et al. (2021), Ljung (2011), Ruitenbeek et al. (2022) and Van Sterkenburg (2019), while the following books were also formative in our thinking about taboo language: Andersson and Trudgill (1990); Jay (1992, 2000); McEnery (2006); Montagu (1967); Pinker (2007). To inform us on the values of attributes, we also scrutinised the tags and definitions in GSW (2007) and Van Sterkenburg (2001), in order to create curated lists of possible values (see 3.2).

Some features of taboo words that are relevant to this project, include the following:

[1] Ethical clearance for the research project was obtained through the Language Matters Ethics Committee of the North-West University (ethics number: NWU-00632-19-A7).

- **Morphosyntactic type**: Taboo constructions include linguistic material on various morphosyntactic levels of independence and compositionality; these types are implemented in TaboeLex as an element `<headwordType>`. In addition to words, it also provides for sub-word items (like affixes), reduced forms (like initialisms), and multiword expressions (MWEs) (see 3.1 for values and examples).

- **Taboo domain**: Much works has been done to identify and delineate the source or reference domains of taboo language, such as religion, sex, scatology, animals, death, disease, etc. Within the scope of this paper, suffice to note that a taboo ontology will be declared as part of the `<denotatum>` element, which is a child element of the `<sense>` element (see 3.2).

- **Taboo type**: While the literal vs. figurative meaning requirement for taboo constructions are still being debated, we take the stance that both constructions with literal meanings, and constructions with figurative meanings could be taboo. For example, while neutral, scientific terms (i.e., orthophemisms) like *penis* and *vagina* could be considered by most people in most contexts as non-taboo, they could still be offensive to some people in some contexts, e.g., they might be dysphemistic in front of one's grandparents at a Christmas dinner, or in a geography class for grade 5 learners.

  This adds a layer of complexity to the development of LDTLs, since homonymous and polysemous constructions need to be handled appropriately. For example, *emmer* refers mostly to 'bucket' (container) – see for example the abridged Dutch dictionary, and the multilingual dictionaries in VDO (2021). However, in some rather obscure cases *emmer* could also refer to 'an inferior person, specifically a prostitute' (i.e., as an abusive term), or 'female genitalia' (i.e., as an obscenity), as reflected in the more comprehensive, unabridged Dikke Van Dale (DVD Online, 2022). This feature of taboo language is practically resolved by introducing the element `<tabooType>` that can be added to any sense of an entry (see 3.2).

- **Tabooness**: Tabooness ratings of constructions will differ between different social groups and are subject to change over time. It is therefore not only essential that constructions should be rated in terms of their observed tabooness in or for certain groups, but also that such ratings should be re-evaluated regularly. For example, it is the task of the British public regulator for communication services, Ofcom, to determine public attitudes towards offensive language on TV and radio, specifically when children are particularly likely to be listening (roughly speaking between 06:00 and 19:00) (Ipsos MORI, 2021a: 3). To this effect, they commission research reports roundabout every five years (Ipsos MORI, 2016, 2021b; Synovate UK, 2010; The Fuse Group, 2005) to determine which words are to be considered mild, moderate, or strong (Ipsos

MORI, 2021a: 4). Similar (but not necessarily comparable) investigations have been done for Dutch in 1998, 2001, 2007, and 2018 (Van Sterkenburg, 2001, 2008, 2019). The element <tabooValue> will capture this knowledge with attribute values on a scale ranging from highlyTaboo to notTaboo; see section 3.2 for other potential values.

- **Context dependence**: Whether a construction is taboo or not, is not only dependent on the situational and/or textual contexts (e.g., whether the derogatory meanings of *emmer* are activated or not), but also on the social context. The word *rambam* ('undefined, imaginary illness') appears only in taboo constructions, like *krijg de rambam* ('get an illness'), but is not considered taboo in most social contexts. The prototypicality rating (<tabooPrototypicality>) will – to a large extent – account for situational, textual, and social contextual dependence of taboo constructions. Words that are taboo in all contexts (e.g., *oetlul* 'jerk, wanker') will get the value alwaysTaboo, while words that are rarely used in the taboo sense (like *emmer*), will have the value rarelyTaboo – see 3.2 for other potential values.

- **Intention and effects**: From a sociopragmatic point of view, taboo language is often defined as language with an expressive/emotive function (Jay, 2020: 39). Hirsch (1985) therefore made a strong case that a taxonomy of taboo language should be based first and foremost on the speech acts (Austin, 1962; Searle, 1969, 1979) in which expressions occur. Following this general approach, we therefore provide for three pragmatic-specific elements, viz. <speechAct> for the type of speech act, <illocution> for the speaker's intention, and <perlocution> for the effect on the hearer (see 3.2).

2.2 Lexical databases of taboo language

We define LDTLs as digital, structured, enriched collections of linguistic constructions that are potentially offensive to some users in some contexts (e.g., in children's books). When implemented in NLP systems as simple look-up lists (gazetteers) for filtering of results, they might sometimes also be called *blacklists*, *greylists*, *swearword stop lists*, or *profanity filters* (e.g., Shutterstock, 2020). Two prominent examples of LDTLs are the following:

- Hurtlex is a lexicon of 1,156 Italian "hate words" that were "linked to synset-based computational lexical resources such as MultiWordNet and BabelNet" (Bassignana et al., 2018).

- Taboo Wordnet is an online, synset-based Japanese resource that could "help detection systems regulate and curb the use of offensive words online" (Choo & Bond, 2021). It consists of 2,095 words with 912 synsets, and it is linked to the Open Multilingual Wordnet.

Besides proprietary lists that are not accessible in the open-data domain, there are also numerous data sets for various taboo-related domains available (see Nakov et al., 2021; Rosenthal et al., 2020; Wiegand et al., 2021; Wiegand et al., 2019; Wiegand et al., 2018; Zampieri et al., 2019b; Zampieri et al., 2020 for overviews of available material). The different tagging schemas of more than 60 such data sets have been compared by Lewandowska-Tomaszczyk et al. (2021), with the aim to create an ontology basis for offensive language identification, while also getting insight in how the concept *offensive* is understood across different projects. They use the term *offensive language* similar to how we use *taboo language* (see 2.1) as a superordinate term for all kinds of language phenomena (Lewandowska-Tomaszczyk et al., 2021: 7). Their proposed ontology of offensive language, together with their methodology for the detection of such language, hold the potential to play an important standardisation role with regards to the treatment of taboo language in the context of Linguistic Linked Open Data (LLOD). In the next phase of our project, their ontology will therefore be the first point of reference to which we will compare our own ontology.

Of utmost importance is that re-usability should be a compulsory design requirement of any LTDL. To make the data re-usable for multiple purposes in several different applications, the database should ideally be rich with as much information as possible – either in the database itself, or otherwise through links to other existing resources. By using subsets of data, or a selection of elements, attributes and/or values, the data could be used in a variety of practical NLP applications like some of the following:

- Offensive language identification (Zampieri et al., 2020) has been a prevailing topic in NLP for a number of years, especially with a view on hate speech, cyber-bullying and abuse detection on social media platforms (Akiwowo et al., 2020; Davidson et al., 2017; Fišer et al., 2018; Jarquín-Vásquez et al., 2020; Korotkova & Chung, 2023; Li et al., 2023; Mostafazadeh Davani et al., 2021; Nakov et al., 2021; Narang et al., 2022; Pradhan et al., 2020; Roberts et al., 2019; Rosenthal et al., 2020; Schmidt & Wiegand, 2017; Teh et al., 2018; Waseem et al., 2017; Zampieri et al., 2019a). The identification of taboo language is also an important aspect of sentiment analysis (Byrne & Corney, 2014; Cachola et al., 2018), especially since the speech acts and language associated with sentiment analysis can oftentimes be more subtle or indirect, e.g., by using humour (Ahuja, 2019; Ahuja et al., 2018; Bansal et al., 2020; Meaney et al., 2021), or irony and sarcasm (Frenda et al., 2022; Husain & Uzuner, 2021).

- More recently the evaluation of large language models for biased and toxic language (Osoba & Welser IV, 2017; Schäfer, 2023; Wiegand et al., 2019) have been pushed to the fore with the public availability of OpenAI's GPT-4 and ChatGPT models. However, from a linguistic and user interface design perspective, our understanding of the implementation of these models in conversational artificial intelligent agents (e.g., speech assistants and chatbots), and especially the relation with taboo language, is still in its infancy.

- LDTLs have been used for many years in applications of text filtering; see Zhou (2019) for an elaborate evaluation of some of these, as well as his own improved implementation. These include, inter alia:
    - **predictive text filtering**, e.g., for search engines, keyboards on mobile phones, online text editors, etc.;
    - **suggestion filtering**, e.g., for spelling checkers and electronic dictionaries (especially dictionary apps for children) that should not suggest swearwords as corrections for ordinary typos;
    - **taboo language censoring**, i.e., redacting, modifying, replacing or removing a word in a text that matches a word in the LDTL; implemented typically as part of parental control software for text, audio, and video (see Porutiu (2023) for an overview and marketing reviews of a number of these applications);
    - **content filtering**, e.g., social media algorithms that (semi-)automatically delete posts or ban users, like Facebook's profanity filter for Facebook Page, or spam filters used in email applications. Other examples of content filtering include e-lexicography tools for choosing good dictionary examples (Kilgarriff et al., 2008), or computer-assisted language learning systems that automatically selects suitable texts for learners (Belaid, 2016).

2.3 Dutch resources of taboo language

Dutch has a rather long tradition in taboo language research, going back to at least 1834 with an history-focused article by J.F. Willems titled *On some old Dutch curses, oaths and exclamations* [translated – the authors] (Willems, 1834). However, the first specialised printed dictionary focusing on language from a taboo domain only appeared in 1977 (EW, 1977). Since then, at least seventeen other printed dictionaries (or dictionary-like books) on various aspects of taboo language have been published (DBG, 1991/2021; GSW, 2007; GT, 1997; HEW, 1988; KDV, 1998; LNS, 1989; LOS, 1990; Lutz-van Elburg, 1990; Lutz-van Elburg & Jager, 1989; NSW, 1984; Van der Gucht et al., 2018; Van der Meulen et al., 2018; Van Lichtenvoorde & Van Lichtenvoorde, 1993; Van Sterkenburg, 2001; WAON, 2013; WEPCT, 2001; WPTG, 2020-2023). Of these, only three are available as digital data: GSW (2007); Van Sterkenburg (2001); WPTG (2020-2023). Since WPTG (2020-2023) is a general dictionary of slang, and therefore also contains many non-taboo constructions, we only use data from the other two dictionaries as candidate taboo constructions for TaboeLex.

One of the most prominent or most used look-up lists of Dutch taboo words (so to see), is the Dutch version of the *List-of-Dirty-Naughty-Obscene-and-Otherwise-Bad-Words*

(RolfBly, 2020).[2] This list was derived from the Dutch section of *The Alternative Dictionaries* (TAD, 2004), although it is not clear when this was done, and by whom it was done. RolfBly (2020) consists of 190 constructions: 165 one-word constructions, and 25 MWEs. While this list will be used in a next phase of the project as one of the baselines for evaluation, several potential problems with the list could already be identified:

- The list is not free of linguistic errors. These include:
    - four spelling errors (i.e., **johny* > *johnny*; **pijpbekkieg* > *pijpbekkie*; **tongzoeng* > *tongzoen*; **triootjeg* > *triootje*);
    - six errors related to obsolete orthographic forms due to spelling reforms in Dutch (i.e., **boerelul* > *boerenlul*; **bokkelul* > *bokkenlul*; **krentekakker* > *krentenkakker*; **kuttelikkertje* > *kutlikkertje*; **paardekop* > *paardenkop*; **paardelul* > *paardenlul*);
    - one compound that should be written as one word (i.e., **trottoir prostituée* > *trottoirprostituée*);
    - an ephemeral word that only exists in TAD (2004) and its derivatives (i.e., *hoempert*, apparently meaning 'hard excrement').
- The list contains only lemmas, e.g., *op+sodemieter · en* (up+tumble · INF 'to fuck off'), and no other word forms, e.g., *op+ge · sodemieter · d* (PTCP). This is particularly problematic for purposes of look-up lists in applications using predictive text filtering, and suggestion filtering (see 2.2). In such applications, the input text cannot be lemmatised first, since filtering needs to happen in real-time and on the fly.
- The MWEs are only presented as lemmas, e.g., *op z'n sodemieter gev · en* (on his carcass give · INF 'to beat the hell out of him'). There is therefore no indication of:
    - orthographic variants, e.g., related to the example above, *zijn/zn/zun* instead of *z'n*, the latter of which does not appear in the 5.9-billion-word nlTenTen20 corpus (Sketch Engine, 2020);
    - morphosyntactic variants, e.g., again related to the above example, *op zijn* (3SG.M) *sodemieter* accounts for only roughly half the cases in the nlTenTen20 corpus; *zijn* is followed by *hun* (3PL), *mijn* (1SG), *ons* (1PL), *de* (DET), and *her* (3SG.F);

[2] An older version (2014) of the list is available at https://github.com/chucknorris-io/swear-words/blob/master/nl, while the list is also reproduced elsewhere on the web.

- lexical variants, e.g., *krijg · en* ('to get') occurs more frequently than *gev · en* ('to give') on the righthand side of *sodemieter* in the nlTenTen20 corpus (Sketch Engine, 2020); or

- syntactic variants, e.g., *geeft hem op zijn sodemieter* instead of *hem op zijn sodemieter geeft.*

- In addition, the MWEs are not always presented uniformly. Compare for instance the lemma *op z'n sodemieter geven* that is presented as a prepositional phrase [$_{\text{PP}}$ *op*$_{\text{PREP}}$ *z'n*$_{\text{PN}}$ *sodemieter*$_{\text{N}}$], followed by the verb [*geven*$_{\text{V}}$]. However, the lemma *reet trappen, voor zijn* has the same [PP V] structure as the former example (i.e., [$_{\text{PP}}$ *voor*$_{\text{PREP}}$ *zijn*$_{\text{PN}}$ *reet*$_{\text{N}}$] [*trappen*$_{\text{V}}$]), but is presented here as [*reet*$_{\text{N}}$ *trappen*$_{\text{V}}$, *voor*$_{\text{PREP}}$ *zijn*$_{\text{PN}}$]. Also, in most cases in the list, only bare verbs are added as lemmas, e.g., *bedonderen* or *belazeren* (both meaning 'to swindle, take someone for a ride'). However, in the case of [*besodemieteren*$_{\text{V}}$] (also meaning 'to swindle, take someone for a ride') a copula verb phrase [*besodemieterd*$_{\text{PTCP}}$ *zijn*$_{\text{COP}}$] ('to have been swindled, taken for a ride) is provided additionally as a separate lemma.

- Numerous polysemous constructions that are most frequently used in a non-taboo way, are included in the list. Compare for instance *achter het raam zitten,* which is an ordinary phrase for 'to sit in a window (looking at what's happening outside)'. However, it is also rarely used with the meaning 'to work as a prostitute' (TAD, 2004), or 'to present oneself in a prostitute-like manner' (DVD Online, 2022). Also compare *welzijn · s+mafia* (welfare · LK+mafia 'ineffective and meddlesome social workers corps') in the list, which is always used unmarked in the Dutch mainstream media.

- Many of the examples are general slang that is not taboo at all. Compare for instance *buffelen* ('to hit; to work hard; to wolf down food'), *huisdealer* ('drugs dealer associated with a certain establishment'), or *kanen* ('to eat'; associated with slang in The Hague).

- Many others are euphemisms, like *de hond uitlaten* ('to let the dog out'), but which can also be used as a euphemism for 'to urinate'. Another example is *de koffer induiken* ('to jump in one's bed'), which is mostly used euphemistically with the meaning 'to have sex'.

- Numerous expected candidates, i.e., highly frequent, highly taboo constructions, are not included in the list. These include words like *debiel* ('mentally deficient'), *trut* ('twat, cunt'), *kanker+wijf* (cancer+woman 'stupid bitch'), and many racial slurs. The list also excludes many English taboo words that are used frequently in Dutch, like *bitch*, *fuck*, and *bullshit.*

A much better and unproblematic list is the *GRoninger OFfensive Lexicon* (GrofLex) (Van der Veen, 2020), a Dutch lexicon of abusive lemmas based on version 1.2 of the Dutch section of HurtLex (Basile, 2020) (see below for more details on Hurtlex). It consists of 847 one-word constructions only (no MWEs). The list has been annotated with part-of speech information, as well as the offensive category (what we call *denotatum* – see 3 below) of each lemma (e.g., ethnic slurs, physical disabilities and diversity, words related to religion, male genitalia, etc.). While the list still contains polysemous constructions (like *kuiken* 'chicken'; *kalf* 'calf'; *druif* 'grape'), and orthophemisms (like *pretentieus* 'pretentious', *fascistisch* 'fascist', *snob* id.), it could be used fruitfully in a next phase of the project as another baseline for evaluation.

3. Design of the TaboeLex lexical database

Our goal is to design an LDTL for Dutch, of which the data can be integrated into various NLP applications and tools, but which can potentially also be useful for human users, or for linguistic research. The general principles and structure of TaboeLex is in line with most existing standards and encoding formats such as Ontolex-Lemon (Cimiano et al., 2016), DMLex (Měchura et al., 2023), LMF,[3] and TEI Lex-0 (Tasovac et al., 2018). General aspects are briefly discussed in section 3.1, followed by those aspects that relates specifically to a LDTL in section 3.2. Figure 1 presents an illustrative example, with LDTL-specific information marked in red. The complete XML schema and documentation, plus eventually all the TaboeLex data, will be made available under a CC BY-SA 4.0 license.

[3] https://www.iso.org/standard/68516.html

```
<lexicographicResource title="TaboeLex" language="ndl">
    <entry id="debiel-word-n">
        <headword>debiel</headword>
        <headwordType>word</headwordType>
        <partOfSpeech tag="noun" />
        <variantForm>dubiel</variantForm>
        <patternForm />
        <linkExternal gigantMolex="12324" />
        <sense>
            <denotatum>entity [person] [mental ability/health]</denotatum>
            <definition language="eng">mentally deficient person</definition>
            <example>
                <text>Mensen laat je toch niet zo opnaaien door die achterlijke
                    debiel.</text>
                <source>nlTenTen20-23694165</source>
            </example>
            <tabooType value="dysphemism">epithet</tabooType>
            <tabooValue value="highlyTaboo"></tabooValue>
            <tabooPrototypicality value="alwaysTaboo"></tabooPrototypicality>
            <speechAct>
                <member value="insult">
                <member value="name-calling">
                <member value="abuse">
            </speechAct>
            <illocution>
                <member value="anger">
                <member value="disrespect">
                <member value="contempt">
            </illocution>
            <perlocution>
                <member value="offensive">
                <member value="derogatory">
                <member value="insulting">
            </perlocution>
            <relation type="synonym">
                <member idref="debiel-word-n" />
                <member idref="idioot-word-n" />
            </relation>
        </sense>
    </entry>
```

Figure 1: Sample entry for *debiel* ('retard; retarded')

3.1 General design

Following our definition of constructions as form-meaning pairings, each taboo construction in the database is defined by aspects related to form, and aspects related to meaning. Regarding form, we use common elements like `<headword>`, `<headwordType>`, `<partOfSpeech>` (of the headword), and `<variantForm>` (e.g., for variants like *f*ck*, *f@ck*, *fark*, etc. for the English loanword *fuck*). The element `<headwordType>` could be extended in future to provide more detailed subcategories, but currently has the following primary values (with Dutch examples):

- `subword`: for affixes (e.g., *· erik* in *bang · erik* (scared · NMLZ 'coward')), and affixoids (e.g., *kanker÷* 'cancer' used as an intensifier in *kanker÷homo* 'bad gay man')[4];

- `reductionForm`: for initialisms like *WTF*;

- `word`: for the uninflected form of words, e.g., *neuk · en* (fuck · INF 'to fuck'); and

- `MWE`: for multiword expressions like:

    - word groups, e.g., *kwark blaffen* ('to ejaculate (male)'), where neither *kwark* ('curd'), nor *blaffen* ('to bark') is taboo, but their combination in a word group is;

    - construction idiom, e.g., *krijg X* ('get X'), used as an imprecation, where X can be various illnesses; and

    - fixed expression, e.g., *Ik kan kakken en pissen en u gemakkelijk missen* ('I can shit and piss without missing you at all').

The rationale behind the element <patternForm> is to include some kind of pattern representation for each headword: on the one hand to allow for the automatic identification of the headword in corpus data (cf. Gantar & Krek, 2022; Odijk, to appear), and on the other hand to deal with the flexibility and variation that many MWEs exhibit. For single words (see Figure 1), the pattern representation is the same as <headword>. For verbal MWEs, the pattern representation is a finite sentence, similar to the way in which patterns are being described in the Corpus Pattern Analysis approach of Hanks (2013). However, rather than using semantic types in the argument slots, we use dummies such as *iemand* 'someone', and *iets* 'something'. See also the recently compiled DUCAME[5] (*DUtch CAnonicalised Multiword Expressions*) resource, and the pattern descriptions in the project *Woordcombinaties*[6].

The last aspect related to the form of an entry, involves the representation of all related word forms of a lemma, e.g., the verb *neuk · en* ('to fuck') has the grammatical forms *neuk* (1SG), *neuk · t* (2/3SG), *neuk · te* (SG.PST), *neuk · ten* (PL.PST), and *ge · neuk · t* (PTCP). Moreover, a comprehensive LDTL should ideally not only include grammatical forms, but also compounds (like *vuist+neuk · en* (fist+fuck · INF 'to fist fuck')), and derivations (like *neuk · er* 'fucker'). This morphological information will be resolved in TaboeLex by means of links (<linkExternal>) to another lexical database, viz.

[4] We use the following notations: middle dot (·) for affix boundaries; divide symbol (÷) for affixoid boundaries; plus symbol (+) for compound boundaries.

[5] https://surfdrive.surf.nl/files/index.php/s/2Maw8O0QTPH0oBP

[6] https://woordcombinaties.ivdnt.org

GiGaNT-Molex[7], the modern part of the computational lexicon of the Dutch language, compiled by the Dutch Language Institute. Because it is linked to GiGaNT-Molex, the full inflectional paradigms and word-formation families of the headwords need not be stored in TaboeLex itself. Instead, this information can be retrieved dynamically from GiGaNT-Molex, if required. This also pertains to MWEs, which are included in GiGaNT-Molex as a whole, and with individual components linked to the appropriate lemmas. This element could also be used in future to link TaboeLex data to other resources, such as thesauri, translation dictionaries, etc.

All information related to the meaning side of a construction are accommodated under the `<sense>` element. While most of its children elements are taboo-specific (see 3.2), three common elements are included, viz. `<definition>` (in English); `<example>`, including the `<text>` and reference to the `<source>`; and `<relation>` to represent lexical relations like synonyms and antonyms.

3.2 LDTL-specific design feature

Various elements, attributes, and/or values that are specific to LDTLs have been added to the design. These are all part of the `<sense>` element since their values can vary depending on which sense of the word is involved; see the information in red in Figure 1. The taboo-specific elements are the following:[8]

- `<denotatum>`: The denotata on a superordinate level are: `event; relation; state; entity; locale; process`. Subtypes provide for constructions related to specific domains; for example, the exonymic epithet *kaas+kop* (cheese+head 'Dutch person') will have the value `entity [person] [inhabitant, citizen]`, while a euphemistic verb like *drukk · en* (press · INF 'to defecate') will be `process [body] [substance] [excretion]`.

- `<tabooType>`: We distinguish four main taboo types on lexicopragmatic grounds, viz.:

  - `orthophemism` (e.g., *penis*);

  - `euphemism` (e.g., *klok-en-hamer-spel* clock-and-hammer-game 'penis');

  - `dysphemism` (e.g., *paal* pole 'penis'); and

[7] https://ivdnt.org/corpora-lexica/gigant/

[8] Since it is impossible in terms of space restrictions to list all possible values for all elements or attributes here, these will be made available as part of the XML schema and documentation; suffice to present here some illustrative examples.