

## Lexical Gaps

Maarten Janssen

Universitat Pompeu Fabra

Maarten.Janssen@upf.edu

Word count: 1711 + 71

### 1. Introduction

When talking about words, linguists mostly focus on those words that are an established part of the vocabulary. However, in some cases it is useful to refer to the words that are *not* part of the vocabulary: the non-existing words. Instead of referring to non-existing words, it is common to speak about *lexical gaps*, since the non-existing words are indications of “holes” in the lexicon of the language that could be filled.

Lexical gaps have never been a main topic of research, but the notion of a non-existing word relates to so many different topics that over the years a significant amount of work has been published on lexical gaps. Since these publications come from a large array of different linguistic fields, the terminology is not always well established. This article will first present a brief overview of the different types of lexical gaps and then discuss some linguistic issues related to lexical gaps.

### 2. A Topology of Lexical Gaps

The study of lexical gaps starts with the work by Chomsky (1965) and Chomsky & Halle (1965). They distinguish between on the one hand *accidental gaps*, which are words that do not exist but could be reasonably expected to exist, and on the other hand *systematic gaps*, which are words that are not even expected to exist since they violate the rules of what a “good” word is. In much of the subsequent work, however, the term *lexical gap* is reserved only for the accidental gaps.

The accidental gaps in the work of Chomsky and Halle are segments or strings of letters that could possibly form words. Such gaps will here be called *formal gaps*, sometimes also referred to as *morpheme gaps*. A significant part of the more recent work on lexical gaps, however, deals with *semantic gaps*. A semantic gap is, in the words of Lehrer (1974), “*the lack of a convenient word to express what (the speaker) wants to speak about.*”<sup>1</sup>, although also words that are possible but not (yet) convenient are considered semantic gaps. A semantic gap is a notion for which there is no word, whereas formal gaps are “words” that do not refer (to any notion). As with formal gaps, we can in principle distinguish between semantic gaps that are accidental, and semantic notion for which no word can exist because they violate the rules of what a “good” notion (for lexicalization) is.

	Accidental	Systematic
Formal	Formal gap	Impossible lexical entry
Semantic	Semantic gap	Non-lexicalizable notion

Table 1. A coarse taxonomy of non-existing words

The class of non-existing words can hence be divided into four different classes as shown in table 1, where a *lexical gap* is typically reserved for the left column only: either a formal or a semantic accidental gap. The remainder of this section provides a more detailed description of these two types of lexical gaps.

#### 2.1 Formal gaps

The orthographic rules and the vocabulary of a language can be used to separate between lexical (occurring) words, possible words, and impossible words. For instance, in English, *apple* is a (lexical) word of the language, *drapple* is a possible word that does not exist, and *drrpple* is an impossible word

---

<sup>1</sup> The term *semantic gap* was introduced by Chomsky, but he did not use it in the same sense; the idea of a semantic gap should more be attributed to Morgan (1968), who called them *possible lexical items*, or to Lehrer (1974), who calls them *functional gaps*.

in English. The word *drapple* is possible because it is a sequence of letters that correctly represents a well-formed phonological sequence in the language. Since different words can be pronounced the same (homophony) or spelled the same (homography), it is useful to distinguish between possible orthographic words, and possible phonological words. Or in terms of the gaps in the lexicon, we can distinguish between (accidental) orthographic gaps and phonological gaps.

Formal gaps can be counted. For instance, by looking at the orthographic gaps: if we consider that words have a finite length and all words consist of combinations of the 26 letters of the alphabet, there is a vast amount of possible sequences of letters, but still a finite number. The possible words form a (very) small subclass of the possible strings, and the lexical words form a small part of the possible words. Formal gaps can be searched for, for instance in the study of neologism: formal neologisms are words that until recently were lexical gaps and have recently been “promoted” to lexical words<sup>2</sup>.

A special type of formal gaps is found in those possible words that can be correctly formed out of lexical words by means of morphological rules. Such words are also called the *potential words*. Many potential words are not merely potential, but also used, and hence lexical words. Those potential words that are not (yet) used are called *morphological gaps*. For instance, from the verb *derive* we can form the lexical word *derivation*, but there is no word *\*derival*. Morphological gaps in the derivational morphology are often called *derivational gaps*, whereas gaps in the inflectional morphology are mostly called *paradigm gaps*. The word *\*derival* is a potential word that is blocked by the existence of *derivation* and is therefore really a lexical gap. However, there is a large group of potential words that are not used, but also not blocked. Take for instance the word *?stratifiability*, which is not in use (at least it does not occur in any English dictionary), but is nevertheless perfectly acceptable and interpretable. For search words, it is unclear whether they should be considered lexical words or derivational gaps.

## 2.2 Semantic gaps

Semantic gaps are those notions for which we have no word to express it. Most instances of semantic lexical gaps are not particularly interesting. To use an example from the Portuguese comedians *Gato Fedorente*: there is no word for the specific type of irritation you feel when you open the fridge to get some milk, and you find that there is no milk. However, there also hardly a reason why such a term should exist, and linguistically, such semantic gaps are hardly of interest.

Semantic gaps typically become of interest when comparing the semantic gaps with lexical words by sketching a matrix of existing word where not all cells of the matrix are filled by a word, therefore also called *matrix gaps*. There is a good amount of work on what type of constraints are responsible for such matrix gaps. A good overview of such constraints is given in Proost (2007).

A specific type of matrix gap is one that is expected to exist in a hierarchy, either a taxonomic or a meronymic hierarchy, but does not exist (see for instance Cruse 2004). An example is the word *dedo* in section 3.2. In the construction of a hierarchy, such gaps often get filled by made-up words, which Fellbaum (1986) calls *pseudo-words*, which are not necessarily word, but just tags to refer to the semantic gap. For instance, he postulates the tag “CREATION-FROM-RAW-MATERIAL” as a way to link a group of verbs, including *weave* and *mold*, that are taxonomically related but have no common hyperonym.

Another type of lexical gap of special interest concerns those notions that are lexicalized in one language, but not in another. For instance, there is no direct translation for the English word *finger* in Spanish (Janssen 2002): there is only the word *dedo*, which is either a finger or a toe. The English word *finger* is therefore an untranslatable word in Spanish, and corresponds to a translational gap. Translational gaps will be discussed in more detail in section 3.2.

An overview of the different types of lexical gaps distinguished in this article is given in table 2, with an indication of how to refer to the non-existing words that “fill” of those gaps.

---

<sup>2</sup> Formal neologisms are by definition words that did not exist before, and they are necessarily possible words, because otherwise they could not have become a word.

Morpheme gap	A sequence of segments that is permitted by phonological rules but not found. Fillers: <i>possible words</i>
Morphological gap	A word that can be generated from an existing word by productive morphological rules. Mostly understood as derivational rules, and therefore also called <i>derivational gaps</i> . Fillers: <i>potential words</i>
Paradigm gap	A morphological gap in the inflectional morphology.
Semantic/functional gap	A lack of a word to express what a speaker might want to talk about.
Taxonomic gap	A gap in the taxonomic structure. Fillers: <i>pseudo-words</i>
Translational gap	A word in one language for which no lexical unit exists in another that expresses that same meaning. Fillers <sup>3</sup> : <i>untranslatable words</i>

Table 2. An overview of types of lexical gaps

### 3. Issues with Lexical Gaps

There are many questions and issues related to lexical gaps, since lexical gaps play a role in so many aspects of linguistics. In this section, two such issues will be discussed.

#### 3.1. Unadapted loanwords

Unadapted loanwords are strange creatures: they are impossible words that are nevertheless lexical words. As already observed by Lehrer (1974, pp. 95), the incorporation of loanwords into the lexicon can itself change the orthographic rules of the language, making the unadapted loanwords possible words, which in turn creates new orthographic gaps for similar words that become possible but not lexicalized.

One could argue that unadapted loanwords do not actually form part of the lexicon of a language, but remain words of the language they belong to; that *blitzkrieg* is a German word independently of how often it is used in English text. In that case, unadapted loanwords should never be considered neologisms, since neologisms are by definition new lexicalized words. However, there are several problems with that view.

Firstly, foreign words are often not incorporated in another language in their full use: the word *goal* in Dutch can only refer to a goal in football, not to any other use it has in English. Also, some impossible words are not even loanwords but coined in a language where they should not be possible (*flexicurity* was coined in Denmark). Furthermore, although loanwords can be unadapted orthographically, they are always at least partially adapted phonetically. And finally, unadapted loanword can form the base of (native) derivations: the English word *windsurf* forms the basis the agentive noun *windsurfista* (windsurfer) in Portuguese, which is neither a possible word in Portuguese, nor in English. So unadapted loanwords differ in various ways from their original in the source language.

#### 3.2. Linking translational gaps

When talking about untranslatable words, people typically quote examples like the Dutch *gezellig* or the Portuguese *saudade*, which are relatively rare cases of culturally dependent untranslatable concepts. However, the majority of lexical mismatches between languages are more mundane. To give two typical examples: (1) in Spanish, the same word is used for the extremities of your hand (fingers) and the extremities of your foot (toes), and (2) French lacks a specific word for a female foal (a filly). These are cases where the translational gap can also be described as a taxonomic gap: cases where a hyperonym or a hyponym is missing.

Notice that *finger* is in no way fundamentally untranslatable in Spanish: there are several strategies for translating such words (see for instance Baker 1992), and in this particular case, one would use either *dedo* or *dedo del mano* (*dedo* of the hand) as the translation depending on the context. But it is a translational gap since there is no direct, single word expression in Spanish for a word in English.

<sup>3</sup> Untranslatable words are not really the fillers of the translational gaps, but rather their synonyms in the source language.

Relations between hyperonyms and hyponyms form the basis of traditional dictionary definitions: the hyponym can be defined in terms of the genus (the hyperonym) and the *differenciae specficiae*, the differentiating features. Therefore, in order to properly relate the Spanish word *dedo* with the English word *finger*, we should not only specify that they are (possible) translations of each other, but also how they differ from each other: that a *finger* is a *dedo*, but specifically one *of the hand*.

A common features in the issues and concepts surrounding lexical gaps is that they have to do with the words that in a sense *almost* exist: studying lexical gaps is studying the outer limits of the lexicon. Given the central role of the lexicon in language, it is therefore hardly surprising that lexical gaps pop up in a wide array of different topics.

## References

- Baker, M. 1992. *In Other Words: a coursebook on translation*. New York: Routledge.
- Chomsky, Noam. 1965. *Aspects of the Theory of Syntax*. Cambridge: MIT Press.
- Chomsky, Noam and Morris Halle. "Some controversial questions in phonological theory." *Journal of Linguistics*.
- Cruse, A. D. 2004, *Meaning in language*, Oxford: Oxford University Press.
- Fellbaum, Christiane. 1996. "WordNet: ein semantisches Netz als Bedeutungstheorie." In: Bedeutung, Konzepte, Bedeutungskonzepte. In: Grabowski, Harras & Herrmann (eds.) *Bedeutungen, Konzepte, Bedeutungskonzepte. Theorie und Anwendung in Linguistik und Psychologie*. Opladen: Westdeutscher Verlag.
- Janssen, M. 2002. *SIMuLLDA: a Multilingual Lexical Database Application using a Structured Interlingua*. PhD Thesis, Utrecht University.
- Lehrer, A. 1974. *Semantic fields and lexical structures*. Amsterdam: North-Holland Publishing Co.
- Morgan, J. L. 1968. "On the notion 'possible lexical item'." In: Darden et al. (eds.) *Papers from the 4th Regional Meeting, Chicago Linguistic Society*. Chicago: Department of Linguistics, University of Chicago
- Proost, K. 2007. *Conceptual structure in lexical items; the lexicalisation of communication concepts in English, German and Dutch*. New York: John Benjamins.