Crafting a lexicon of referential expressions for NLG applications

Ariel Gutman, Alexandros Andre Chaaraoui, Pascal Fleury
Google Research Europe, Zurich
{relgu, alexandrosc, fleury}@google.com

Abstract

To engage users, a natural language generation system must produce grammatically correct and eloquent sentences. A simple NLG architecture may consist of a template repository coupled with a lexicon containing grammatically-annotated lexical expressions referring to the entities that are present in the domain of the system. The morphosyntactic features associated with these expressions are crucial to render grammatical and natural-sounding sentences. Existing electronic resources, like dictionaries or thesauri, lack wide-scale coverage of such referential expressions. In this work, we focus on the creation of a large-scale lexicon of referential expressions, relying on n-gram models, morpho-syntactic parsing, and non-linguistic knowledge. We describe the collected linguistic information and the techniques used to perform automatic extraction from large text corpora in a way that scales across languages and over millions of entities.

Keywords: Natural Language Generation, Lexicon Extraction, Referential Expressions

1. Introduction

Dialogue systems, such as voice-driven personal assistants or conversational chat-bots, as well as other natural language generation (NLG) applications are bound to produce appropriate, grammatical and well-formulated utterances, in order to engage the human user. One often-overlooked prerequisite for such behaviour is the use of correct lexical information regarding the entities in the domain of the system (e.g., place names, names of people, etc.). In this paper, we shall describe several techniques that make it possible to acquire such information automatically at a large scale.

A typical architecture of an NLG system has distinct modules for content planning, sentence planning and sentence realization, as outlined by Reiter and Dale (2000) or Walker and Rambow (2002). A simple sentence realization module may contain the following two components:

1. A template repository, which stores the various messages which the system can generate. These templates, each created for a specific communicative intent of the system, may correspond broadly speaking to the notion of *constructions* of the *construction grammar* framework (Goldberg, 1995): they are a mixture of lexical, syntactic and surface form specifications for each utterance.

2. The lexicon, containing the lexical forms (lexemes) and the relevant grammatical information of the entities in the domain of the system.

The usage of a template-based sentence realization system is, of course, quite old (see Weber and Mendoza (1973) for a description of a very early system which produces haikus). In their simplest form, template-based systems have been contrasted with true NLG (Reiter, 1995). Yet the addition of the second component, namely a linguistically annotated lexicon, makes them truly NLG-worthy. NLG lexica have typically been hand-crafted, but this is not possible if the scale of the required domain is very big (e.g. weather reports for all localities on Earth).

As stated above, in this paper we are concerned with the automatic crafting of such large-scale lexica in a multilingual setting. Morphosyntax and surface form variations are very language-specific, as will be illustrated below with some languages for which we created lexica: Czech, English, French, Swedish and Russian. We are especially interested in acquiring information about *referential expressions*, i.e. expressions which have specific referents in the world (either real or fictional), e.g. *Paris*, *The Beatles*, or *James Bond*. Such expressions are often termed *proper nouns* or *proper names*; in either case we note that they can superficially seem as compositional noun phrases, such as *The Great Lakes*.

Being noun phrases, these referential expressions exhibit grammatical properties that can affect the selection and form of surrounding words, due to phenomena such as grammatical agreement, preposition selection and the like. Therefore, they cannot simply be plugged into an empty slot in the template, as part of the template may need to be re-edited. Instead, the template needs to be specified in such a way that this lexical information is taken into account. Moreover, in some cases, the combination of information from multiple referential expressions is needed to generate the grammatically correct form of a sentence. This happens, for example, with the gender of a list of conjoined nouns in French: a single masculine noun in it will trigger masculine agreement with any element dependent on the list.

An important property of referential expressions, in contrast to more conventional lexemes of a language, is their large scale. Thus, the Second Edition of the 20-volume Oxford English Dictionary contains about 300,000 entries (Simpson and Weiner, 1989), yet the number of referential expressions is theoretically unlimited and in practice could reach tens of millions, depending on the domain of the NLG system. This immense richness of referential expressions is often overlooked since many NLU systems, such as parsers, do not require grammatical information about these names: it suffices for an NLU system to mark these names as such. If moreover, the referential expression is compositional, its proper name nature can be overlooked.

Thus, most electronic lexical resources concentrate on the common lexemes of language, such as common nouns,

verbs or adjectives. For instance, Sagot (2010) presents a lexical database of French containing about 110,000 lemmas, out of which only about half are proper nouns. Moreover, the grammatical information needed for proper nouns is often not encoded in standard lexical resources or dictionaries. For example, in some languages various toponyms require different locative prepositions (for instance, islands require in general the preposition "on" in English, though some larger islands, or island groups, are exempt). Such information is usually not present in dictionaries, or it can only be deduced from examples given there.

In this paper, we present three different systems to acquire large-scale lexical data consisting mainly of referential expressions (as well as common nouns), in a multilingual setting. Two of the systems use data-mining methods to extract information from corpora, in which referential expressions are marked and linked to an entity's identifier in a non-linguistic knowledge base of entities, such as a geographical repository or a database of people. The corpora we used include Wikipedia pages, as well as selected news sites. The difference between the two approaches is related to the amount of grammatical annotation the corpus has. For some languages, which we call "high-resource languages", a parser may be at our disposition, while for others, called here "low-resource languages" we have no such tools. The third system is a last-resort rule-based system which "guesses" the grammatical properties of a given referential expression using available knowledge at the time of generation.

We present below a simple example of the type of information we want to acquire, and subsequently the three systems.

2. A simple example of a lexicon

Consider an NLG system which produces weather reports for various localities. It may contain a template as the following:

```
It is sunny in (Location).
```

In this template, the placeholder `(Location)` is to be replaced with a name of a location (a *toponym*):

```
It is sunny in Paris.
```

Yet it is easy to see that such a simplistic template would generate ungrammatical sentences if the location requires a different preposition, as is typically the case with islands or lakes:

```
It is sunny on Tenerife.
It is sunny at Lake Como.
```

This last example also illustrates that the possible choices are constrained by the referential expression, but also by the wanted semantics, as *on Lake Como* would be another perfectly acceptable phrase in this context, but with a slightly different meaning.

To accommodate such cases, the template has to be rewritten so that the correct preposition is chosen:

```
It is sunny (Locative
preposition + Location).
```

Once the template has been amended, the system now relies on the correct preposition being specified in the lexicon for each entity (see Table 1).

Name	Preposition
Paris	in
Tenerife	on
Lake Como	at

Table 1: Samples of different locative prepositions in English.

A further complication is presented by toponyms such as *the Isle of Man*, for which we expect the following message:

```
It is sunny in the Isle of Man.
```

Yet the determiner *the* is not an integral part of the toponym, as is evident from the fact that it can be removed in certain expressions (*Britain's Isle of Man*) and would not appear in a listing of countries or on a map. Thus, the lexicon needs to be augmented with information about determiners as shown in Table 2.

Name	Preposition	Determiner
Paris	in	-
Tenerife	on	-
Lake Como	at	-
Isle of Man	on	the

Table 2: Locative prepositions and the required determiner for different English toponym samples.

An English lexicon may additionally contain traditional grammatical information about gender and number, to be used for instance in pronominalization or verbal agreement, or phonological information, such as whether a lexeme starts with a vowel. To exemplify the latter, contrast *Australia* with *Uruguay*, where only the former has a vocalic onset, yielding expressions like *an Australian city* versus *a Uruguayan city*. In languages with richer morphology like Russian, the lexicon may additionally enumerate the various case inflections of a given name, which are often idiosyncratic for proper nouns, or provide other necessary pieces of grammatical information, such as animacy in Russian. Apart from the grammatical information, the lexicon may be enriched with multiple names for a given entity, be it short or long versions of the same name (*Frankfurt* vs. *Frankfurt am Main*) or various nicknames of entities (*the Big Apple* vs. *New York*).

3. N-gram-based lexicon extraction

For low-resource languages, i.e. languages for which some amount of written material can be found in the web, we have at our disposition a corpus of texts lacking grammatical annotation. A prerequisite of the lexicon extraction

process, however, is that the potential referential expressions are identified in the corpus, and are linked to the relevant entities in the knowledge base of the system, a process known as named-entity extraction (Momchev, 2010). Since in this case we do not possess any grammatical annotation of the text, we rely on the insight that functional words in the vicinity of the referential expressions may give us information regarding the grammatical features of the expression, a method that has been shown to explain similar aspects of child language acquisition (Gutman et al., 2015). For instance, if we want to deduce the gender of the French toponym *Paris*, we may observe the presence of the masculine determiner *le* in the expression *le grand Paris* and deduce that Paris is a masculine toponym. At the same time, we may observe the text *Paris est belle*, from which we would deduce that it is actually a feminine toponym, probably due to the feminine gender of the latent concept *ville* ("city"). This hints at the fact that such proper nouns usually do not have a fixed grammatical gender, a property which could potentially also be modeled by the extracted annotations.

In practice, however, in order to use this procedure, we provide for each language only a short table of functional words (typically determiners) associated with their grammatical properties. For example, for French we used the data presented in Table 3. In this table, grammatical features are shown in the columns, the functional words in rows, and the modeled *attributes* in the cells. Note that some function words do not provide any information regarding a given feature, so the corresponding table cell is empty, e.g. the plural determiners that are gender-neutral (or underspecified) in French. Conversely, one form may be associated with competing features: in German, the determiner *die* can be either feminine singular or gender-neutral plural, and the determiner *der* could be masculine singular nominative or feminine singular genitive.

	Gender	Number	Elision
le	masc.	sg.	-
la	fem.	sg.	-
l'		sg.	+
les		pl.	
un	masc.	sg.	
une	fem.	sg.	
des		pl.	

Table 3: Gender, number and whether elision is applied or not for French definite and indefinite articles.

Additional data given to the system is whether these words should appear before or after the corresponding referential expression (French and German determiners appear before), and the size of the n-gram window around the named entity to examine. In practice, looking at bigrams proved to be sufficient. For features like elision-triggering, which is a sandhi phenomenon (i.e., word-edge variation which is due to morpho-phonological conditions), the system only considers the unigram adjacent to the referential expression.

Given this data, the assignment of grammatical features to referential expressions is straightforward: for every mention m of referential expression E in the set of mentions M_E, for each grammatical feature F, and for each possible attribute value a_F of the feature, the system identifies the functional words t in the window Ω_m of n-grams adjacent to the mention of the referential expression. This contributes a certain weight $w_{a_F,t}$ to the total score of the given attribute of the expression $a_{F,E}$. The score is normalized by the number of mentions $|M_E|$.

$$score(a_{F,E}) = \frac{\sum_{m \in M_E, t \in \Omega_m} w_{a_F,t}}{|M_E|} \quad (1)$$

Selection of the right attribute for a given feature F of a referential expression E is then done by taking the highest scoring attribute (in the set of possible attributes A_F), above a certain threshold min_{a_F}:

$$a_{F,E} = \arg\max_{a \in A_F} \{score(a_{F,E}) | score > min_{a_F}\} \quad (2)$$

The confidence threshold min_{a_F} may be used in order to filter out cases where there is not enough supporting evidence for an attribute in the whole corpus. Yet in practice, as we shall see below, setting this threshold to zero allows us getting maximal coverage without compromising the quality of the results significantly.

As for the calculation of the weight $w_{a_F,t}$ this could in principle be learned from an annotated corpus. Yet since we do not have such annotations, we take a simple approach of distributing a weight of 1 over all possible attributes $A_{F,t}$ of a feature F specified for a certain functional word t:

$$w_{a_F,t} = \begin{cases} \frac{1}{|A_{F,t}|} & \text{if } a_F \in A_{F,t} \\ 0 & \text{otherwise} \end{cases} \quad (3)$$

For example, the weight of the attribute *masculine* of the French determiner *le* is 1, while the weight of the same attribute for *les* is 0 (since no gender is specified for *les*). In the experiments we did with French and Swedish there were no cases of fractional weights, since every functional word has at most one attribute specified for each feature.

Using this approach we extracted about 800,000 lexicon entries. We selected a sample of 100 entities to evaluate the precision of the grammatical features of gender, number and elision. The results are given in Table 4, using two different confidence thresholds: 0% (i.e. no threshold) and 10%. These results are compared to a baseline result, which consists of uniformly selecting the majority group (i.e. masculine, singular and no elision). As expected, using a higher threshold increases the precision,[1] though this comes with a decreased coverage of about 40%, compared to the zero-threshold results.[2] The rest of the figures in this paper are given for the case when a zero confidence threshold is used.

[1] Surprisingly, the precision goes slightly down for the number feature. This can probably be ascribed to the usage of a small sample and the very high initial precision rate.

[2] To be more exact, out of the sample of 100 entities, only 58 entities get the gender or elision features assigned with the 10% confidence threshold, and similarly only 72 entities get the number feature assigned.

The referential expressions in the sample are a mixture of proper nouns (e.g. *Dheepan* or *Nathalie Rihouet*), proper names (*Miss France 2007*), acronyms (*FICP = Fichier national des Incidents de remboursement des Crédits aux Particuliers*) as well as common nouns (*neuvaine*) or noun phrases (*perche à selfie*). All refer to entities in the domain of the system and as mentioned before French toponyms or company names do not always have a fixed gender. For this evaluation we relied on the gender as it appears in the French Wiktionary.[3] If no gender was given, we did not include the entity in our evaluation and therefore we did not calculate a recall value.

French	Gender	Number	Elision
Baseline	60%	82%	76%
0% threshold	87%	97%	98%
10% threshold	98%	96%	100%

Table 4: Precision results obtained for French grammatical features applying n-gram based lexicon extraction, with two different confidence thresholds. For comparison, a baseline of selecting the majority group is given as well.

The low score obtained for the gender feature, when no threshold filtering is used, can be explained by the fact that plural articles (as well as the elided article *l'*) neutralize the gender property. For example, the determiners in *l'Autriche* or *les Maldives* do not provide any information about the gender. Yet if our corpus contains a mistyped expression such as *le Maldives* (and such typos are frequent in web corpora), the system will erroneously deduce that *Maldives* is masculine in the lack of counter-evidence. This is rectified to some degree by filtering the results using a minimal scoring threshold, which we did not, however, use in the evaluation procedure. For instance, setting the threshold to 0.1 (i.e. the evidence for gender is present in at least 10% of the occurrences of every given expression) increases the gender precision to 90% while purging 30% of expressions.

The same technique was applied to Swedish, using various Swedish determiners. We used the various forms of the definite article *den*, the indefinite article *en*, the demonstrative *denna*, the possessive pronouns as *min* ("my"), as well as other determiners: *vilken* ("which"), *någon* ("some"), *ingen* ("no"), and *annan* ("another"). All these determiners exhibit number variation as well as gender variation in the singular (common or neuter gender). For Swedish we used a smaller corpus and extracted about 35,000 entities.

The precision results are shown in Table 5, evaluated on a sample of 115 common nouns and 150 proper names. The baseline results are given for an equal mix of proper and common nouns.

Here too, the lower result for gender can be explained by neutralisation of the gender feature in plural determiners. In an expression like *de nya Flugbussarna* ("the new Airport-busses") there is no information regarding the gender of the referential expression *Flugbussarna.*

[3] `http://fr.wiktionary.org`.

Swedish	Gender	Number
Baseline (mixed)	52%	85%
Common nouns	90%	97%
Proper names	66%	92%

Table 5: Precision results obtained for Swedish grammatical features applying n-gram-based lexicon extraction, with no confidence threshold. For comparison, a baseline of selecting the majority group is given as well.

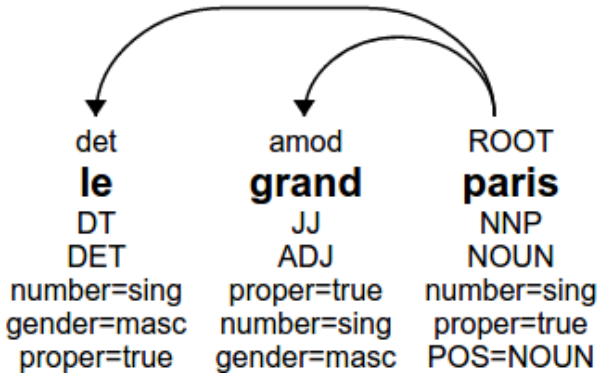

Figure 1: Extracting grammatical properties (number=singular and gender=masculine) from a determiner (DET) and an attributive adjective (ADJ). The labels on the arcs permit the extraction system to find the words which may carry the relevant information (det=determiner arc, amod=attributive modifier arc).

4. Dependency-tree-based lexicon extraction

For languages for which we have access to a morpho-syntactic parser, we use a more involved system. Specifically, the morpho-syntactic parser presented in Andor et al. (2016), annotates our corpora with dependency relations and with some morphological annotations. Occasionally, the referential expression itself is annotated with the desired grammatical features (such as the grammatical gender and number) yet this is not always the case for proper nouns. Essentially, we use the same technique as before, but instead of guessing that a nearby determiner is related to the target expression, we can identify the correct determiner by virtue of the available syntactic parse (following a dependency arc). Moreover, we are not limited to specific functional items, but we can also rely on agreement morphology apparent on verbs or adjectives.

For example, we can extract the gender of *Paris* both from a determiner and an attributive adjective in the phrase *le grand Paris* and from the predicative adjective in the sentence *Paris est belle*, corresponding to the dependency trees shown in Figures 1 and 2.

Note that in both cases the parser does not give us the grammatical gender of the name *Paris*, possibly due to the difficulty of assigning such a gender.

Similarly, we can directly count which prepositions govern each referential expression in order to infer the most common locative preposition. Of course, to infer phonological sandhi features (such as the *elision* feature), the extraction

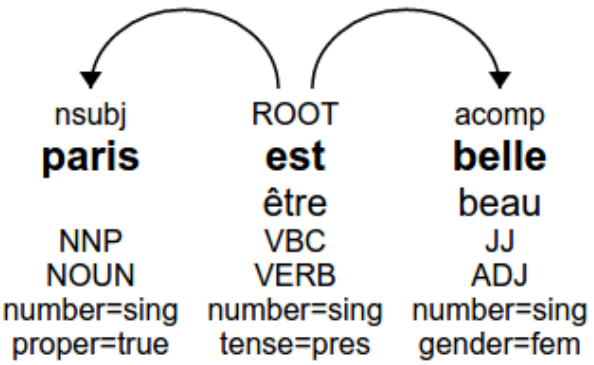

Figure 2: Extracting grammatical properties (number=singular and gender=feminine) from a predicative adjective (ADJ). Here the extraction system follows two arcs: *Paris* is the nominal subject (nsubj) of the verb *est* ("is", being the root of the tree), while *belle* is an attributive complement (acomp) of the verb.

system must still take into consideration linear adjacency rather than dependency relations.

For this process, we used a much larger corpus, and managed to extract about 7 million French lexical entries, being mostly proper names. Thus, for evaluation we used a larger evaluation set, consisting of about 46,000 entries. The precision results are given in Table 6.

French	Gender	Number	Elision
Precision	70%	98%	95%

Table 6: Precision results obtained for French grammatical features applying dependency-tree-based lexicon extraction.

We note that the results are worse than the n-gram based model, especially for the gender feature. This is expected, since we are able to infer such properties also when no article is present (for instance by looking at a predicative adjective, as in Figure 2), but this necessarily increases the noise in the system.

Using this system we have also extracted the locative preposition of toponyms. Here we got a precision level of 88%.

5. Lexicon inference based on minimal information

In some cases our methods of lexicon extraction are not practicable at all, or they failed for a specific entity. Yet we may still have at our disposition *non-linguistic knowledge* about the entity coupled with some default (typically *official*) name (for instance, we may have a database of geographical names or of movie actors). In such cases we can still apply some last-resort rules to guess the relevant grammatical properties, either by detecting some morpho-syntactic pattern in the name itself, and/or by relying on the non-linguistic information.

A trivial case is if a French name starts with an article: in that case we can infer the grammatical properties directly from that article, as in the toponyms *Le Havre* or *La Rochelle.*

A less-trivial example is using the ending of a French name to infer its gender. Our investigation shows that relying on a simple heuristic of assigning feminine gender to French names ending with *-e* is correct in about two thirds of the cases.

As for non-linguistic information, if we know, for instance, that an English geographical name represents an *island*, we can guess with high probability that it should take the locative preposition *on*. Additionally, we can detect the word "island" in the name itself and apply the same heuristic. Similarly, for names of people, we may assume that the gender of the named person corresponds to the grammatical gender of the name.

We have applied this method specifically to a set of approximately 11,000 Czech toponyms, with the goal of obtaining their locative prepositions to form prepositional phrases such as *v Praze* ("in Prague"), *ve Vancouveru* ("in Vancouver"), or *na Ukrajině* ("in Ukraine"). Based on the knowledge base of the system, entities have been classified in different categories that share linguistic properties with regards to the locative preposition: expressions referring to islands, mountains, peninsulas, airports, train stations, highways, universities, castles or lakes, were assigned the locative preposition *na*, while other expressions were assigned the locative preposition *v* or its allomorph *ve*, based on the presence of certain consonantal onsets in the referential expression. Results were evaluated with a golden set of 1,200 manually annotated toponyms, where subsets were chosen based on the entity's frequency in the corpus (see Table 7).

Sample set	Set size	Precision
Head - 1st tertile	400	96%
Torso - 2nd tertile	400	98%
Tail - 3rd tertile	400	99%

Table 7: Precision of locative preposition assignment for Czech toponyms using lexicon inference based on the type and the orthographic name of the entity.

Note that Czech nouns inflect for the locative case after these prepositions. In order to acquire the paradigm of the Czech names we still had to use an n-gram-based lexicon extraction process, in which we could identify case inflections by virtue of their co-occurrence with certain prepositions.

6. Conclusions

In this paper we presented various techniques to assemble information about referential expressions known more generally as *proper names*. We showed that given a corpus with annotation of referential expressions alone, we may use minimal grammatical knowledge of functional words in the language in order to infer grammatical properties. If we do have grammatical annotation we may use these to improve upon the impoverished technique.

Finally, we suggested that even when no linguistic knowledge apart from the name of an entity is available, we may

still rely on that name together with non-linguistic information about the entity to infer some grammatical properties with some confidence. In this respect, as illustrated in Figure 3, the three presented methods can be combined; especially the lexicon inference can serve as a last-resort method to assign linguistic properties to expressions which are only rarely found in the available corpora.[4] Conversely, if certain grammatical properties are generally predictable from the orthography of a name or the entity's type, we may choose to mainly rely on this method and only store in our lexicon the exceptions to the rule (which can be gathered using lexicon extraction).

In future work, we aim to address methods for selecting and grouping various referential expressions referring to the same entity. While in the simplest case we may just select the most frequently occurring referential expression as the relevant one (as we did in the above experiments), the situation is more complicated if we want to reconcile several expressions into a paradigm, as in a case-inflecting language. This can be achieved if we have some minimal knowledge of the relevant paradigms present in the language, similarly to the techniques used by Clément et al. (2004) for French verbs. A further problem is to find several different referential expressions, or paradigms of such, differing in some semantic dimension. For example, one expression could be an *official* name, and another the everyday *colloquial* name. This is in fact quite a difficult task, which warrants a separate discussion.

7. Acknowledgements

We would like to thank Ivan Korotkov, Jana Strnadova and Daniel Calvelo Aros for creating and working on different parts of the above described systems, as well as many other linguists and engineers who contributed to our work.

8. Bibliographical References

Andor, D., Alberti, C., Weiss, D., Severyn, A., Presta, A., Ganchev, K., Petrov, S., and Collins, M. (2016). Globally normalized transition-based neural networks. *CoRR*.

Clément, L., Lang, B., and Sagot, B. (2004). Morphology based automatic acquisition of large-coverage lexica. In *LREC 04*, pages 1841–1844, Lisbonne, Portugal.

Goldberg, A. (1995). *Constructions: A Construction Grammar Approach to Argument Structure*. Cognitive Theory of Language & Culture. University of Chicago Press.

Gutman, A., Dautriche, I., Crabbé, B., and Christophe, A. (2015). Bootstrapping the syntactic bootstrapper: Probabilistic labeling of prosodic phrases. *Language Acquisition*, 22(3):285–309.

Momchev, N. (2010). Annotating web documents with Wikipedia entities. Master's thesis, Sofia University.

Reiter, E. and Dale, R. (2000). *Building Natural Language Generation Systems*. Cambridge University Press.

Reiter, E. (1995). NLG vs. templates. In *Proc of the Fifth European Workshop on Natural-Language Generation (ENLGW-1995)*, Leiden.

Sagot, B. (2010). The Lefff, a freely available and large-coverage morphological and syntactic lexicon for French. In *7th international conference on Language Resources and Evaluation (LREC 2010)*, Valletta, Malta, May.

John Simpson et al., editors. (1989). *The Oxford English Dictionary*. Oxford University Press, Oxford, second edition.

Marilyn Walker et al., editors. (2002). *Computer Speech and Language: Special Issue on Spoken Language Generation*, volume 16(3–4). Academic Press.

R.L. Weber et al., editors. (1973). *A Random Walk in Science*. Institute of Physics Publishing, Bristol and Philadelphia.

[4]Note that for a given language, we typically only use one of the corpus-based lexicon-extraction methods, depending on the availability of a dependency parser for that language. The combination of these two methods is required for the construction of a multi-lingual lexicon.

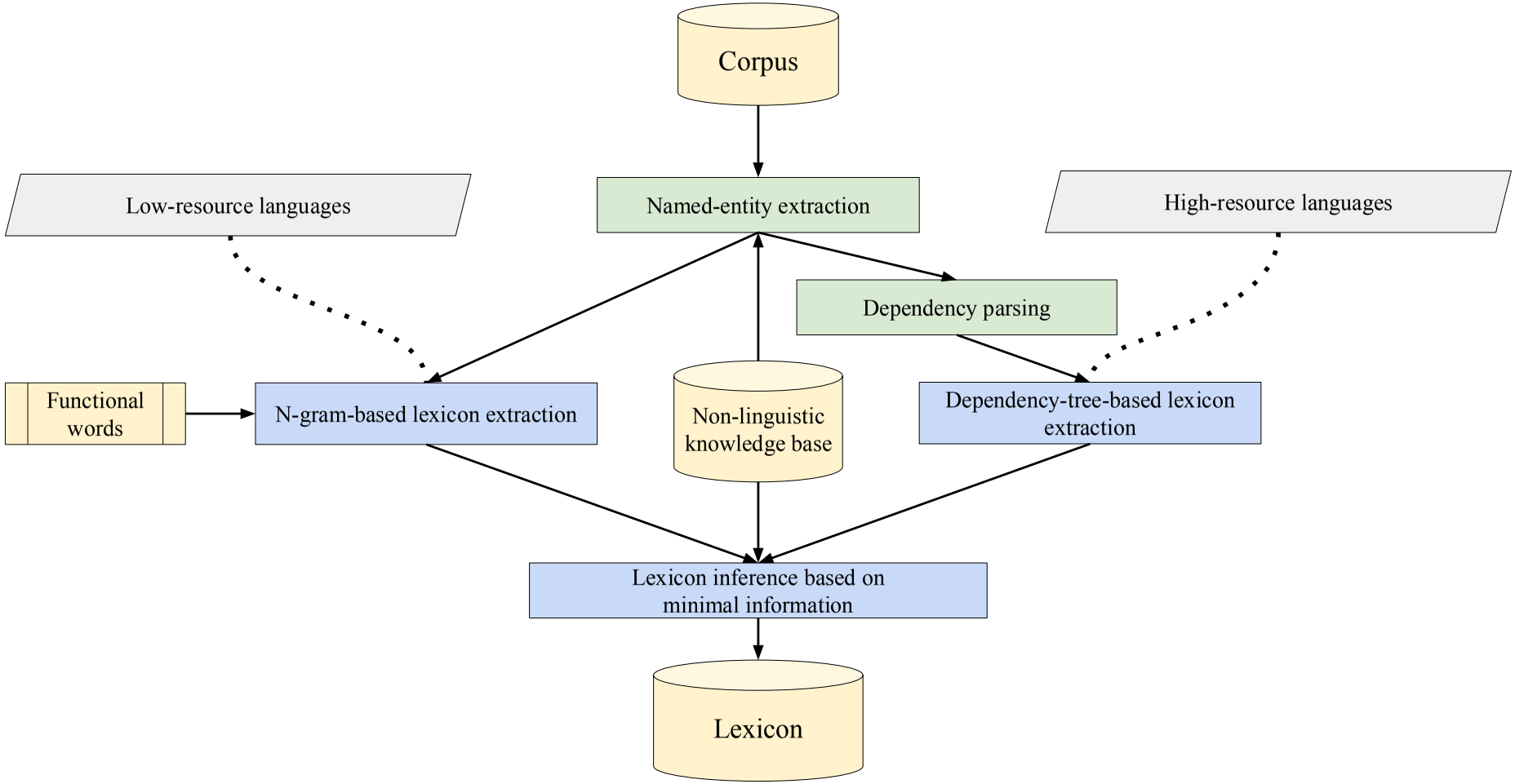

Figure 3: Flow diagram of the proposed architecture for crafting a linguistically annotated lexicon. Note that the lexicon inference based on minimal information is optional.