Automatic acquisition of domain-specific morphological resources from thesauri

Natalia Grabar and Pierre Zweigenbaum
Service d'Informatique Médicale / DSI / Assistance Publique – Hôpitaux de Paris &
Département de Biomathématiques, Université Paris 6, Paris, France
{ngr,pz}@biomath.jussieu.fr http://www.biomath.jussieu.fr/~pz/

Abstract

These is a growing body of evidence that morphological knowledge, beyond stemming methods, can help Information Retrieval tasks. Automated acquisition methods have been devised to learn morphological knowledge from corpora and thesauri, and show how corpus- or domain-specific morphological knowledge can enhance information retrieval results. In medicine, compounding and derivation are commonly used to form new words (*e.g.*, "neo-classical compounds"). However, although there exist many large medical thesauri, medical morphological resources other than for inflection remain scarce in most languages. We therefore designed a method that learns morphological knowledge from an input thesaurus and identifies morphological families of word forms. It can work with no a priori linguistic resources, and has been applied to several languages (French, English and Russian). Applied to a medical thesaurus in these three languages, it produces morphological families with a precision ranging from 91.9% (English) to 99.6% (Russian). We also examined the influence of part-of-speech tagging and lemmatization on the quality of the results.

1 Introduction

There is a growing body of evidence that morphological knowledge, beyond stemming methods, can help Information Retrieval tasks (*e.g.*, (Krovetz, 1993)). However, this depends on the availability and coverage of such resources. Three types of morphological variation are classically considered. *Inflection* produces the various forms of a same word such as plural, feminine or the multiple forms of a verb according to person, tense, etc. Inflectional morphology is well handled in many languages, and *lemmatizers*, which reduce inflected forms to their canonical forms, are available for a large set of languages (for instance for French, Silberztein's INTEX system (Silberztein, 1993) based on LADL's DELAF dictionaries, or Namer's FLEMM lemmatizer (Toussaint *et al.*, 1998)). *Derivation* is used to obtain, *e.g.*, the adjectival form of a noun (noun *vesicule* yields adjective *vesicular*). *Compounding* combines several radicals to obtain complex word forms (*e.g.*, *vesicule* + *virus* yields *vesiculovirus*). The CELEX base describes inflectional and derivational knowledge for Dutch, English and German (Burnage, 1990). However, the general situation for most other languages, including French, is that no complete description of derivational or compositional morphology is available. Derivational morphology is being addressed in recent projects (*e.g.*, the MorTAL project on French derivational morphology (Dal *et al.*, 1999)).

As many technical domains, medical terminology very often calls on compounding and derivation to form new words. Many medical words are formed by compounding of Latin and Greek radicals: the so-called *neo-classical compounds* (Wolff, 1987). For instance, *adren–*, *myel–* and *neur–* are prefixed to *pathy* to form *adrenomyeloneuropathy*. Therefore, medical morphology is acknowledged as an important area of medical language processing and medical information indexing (Wingert *et al.*, 1989; Schulz *et al.*, 1999), and there is a need to develop resources for it. This is the aim of the present work.

A specific inflectional and derivational knowledge base has been prepared for medical English in the

framework of the UMLS project (McCray *et al.*, 1994). Building medical morphological resources for several languages is also beginning to be tackled in a global way (Schulz *et al.*, 1999). The specific goal of the present work is to investigate methods that take advantage of the wealth of existing medical terminologies to help acquire morphological knowledge for these languages (Grabar & Zweigenbaum, 1999; Zweigenbaum & Grabar, 2000a). We aim to learn *morphological families*: connected graphs of word forms, where two forms are linked if they entertain a morphological relation (inflection, derivation, compounding or a combination thereof). Such families can be used, *e.g.*, for query expansion.

Automated acquisition methods have been devised by several researchers to learn morphological knowledge from available resources. The general idea is to obtain constraining contexts in which pairs of word forms that "look alike" actually belong to the same morphological family. "Looking alike" means having similar strings: for instance, sharing a long enough prefix string.

The constraining contexts may be drawn from a large corpus with a variant of mutual information statistics (Xu & Croft, 1998). Two word forms reduced to a common root by an "aggressive" stemmer and that co-occur significantly more than would be expected if they were independent are considered to belong to a same equivalence class. This correctly identifies, *e.g.*, {*bond*, *bonds*} and {*animation*, *animators*} as being related, but also correctly rejects {*policy*, *police*}.

Contrasting thesaurus terms with corpus occurrences of morphologically similar words in a given domain (Jacquemin, 1997) provides another favorable situation. Given a two-word thesaurus term, corpus collocates that consist of two word forms similar to the term words (*i.e.*, they share an initial common substring) have a high probability of being morphological variants of the thesaurus term. For instance that method identifies collocates *gene expression* and *genic expression*, where *gene* and *genic* do belong to the same morphological family.

A focussed study of part-of-speech-tagged word forms is yet another way of identifying related word forms (Dal *et al.*, 1999). Given a tentative suffixation rule such as $V \xrightarrow{+able} A$, one can match verbs to adjectives ending in *-able* and identify those that have a high chance of being derived from one another.

Our method performs a sort of reverse engineering of the linguistic knowledge that a terminologist used to author a thesaurus. Thesauri may specify different kinds of relations between terms and concepts. Given a *concept*, they often include both a preferred term and synonym terms. Synonym terms often call on pairs of morphologically related words: for instance, the two terms *Acute suppuration of nasal sinus* and *Acute suppurative inflammation of nasal sinus*, which are declared synonym in the SNOMED medical terminology (see below, section 1), rely on the pair of morphologically related word forms {*suppuration*, *suppurative*}. The basis of our method is to identify such word pairs in the local context of synonym terms. We then induce morphological rules that we apply to a reference list of word forms to collect a larger set of word pairs, which we join into morphological families.

Various kinds of morphological knowledge can be directly or indirectly useful for information retrieval.

Families of morphologically related word forms (example (1)[1]) relate together series of word forms that share a common stem. A morphological family could provide the basis for an equivalence class for retrieval. The stem (or equivalent full word) could also be used to index each word of the family.

(1) *abdomen, abdominal, abdominalis, abdominis, abdomino, abdominocentesis, abdominopelvic, abdominoplasty (E);*
aorte, aortique, aortite, aorto (F);

[1]When appropriate, we specify the language of the examples with *(F)*, *(E)* or *(R)* for French, English or Russian respectively.

cardia, cardiaque, cardiaques, cardio, cardiomégalie, cardiopathie, cardiopathies, cardite (F);
актиний, актиния, актинический, актиническая (R)

Morphologically related pairs of word forms (example (2)) are the basic building blocks of morphological families.

(2) *abdominal, abdominaux (inflection) (F);*
аденоз, аденозы (inflection) (R);
aorta, aortic (derivation) (E);
аорта, аортальный (derivation) (R);
aorta, aortitis (compounding) (F);
aorta, aortogram (compounding) (E);
фибро, фиброма (compounding) (R);
слизистое, слизистоподобное (compounding) (R)

Morphological rules that relate a word form with another, morphologically related word form (example (3)) are a useful resource to produce the above resources (1) and (2).

(3) *aorta* $\overset{a|itis}{\leftrightarrow}$ *aortitis*
Examples : l|ux, a|ic, a|itis, a|ogram

All the methods presented in this article deal with *strings* rather than *morphemes*. However, we often use words such as *prefix*, *suffix* or *radical* to denote strings that might play the role of such word parts.

In the rest of the paper, we describe our acquisition method and experiments performed with it on a medical thesaurus in French, English and Russian. We first present these thesauri (section 2), then the method and its results (section 3). We also study how these results are affected if part-of-speech-tagged or lemmatized input is available (section 4). We finally synthesize the results (section 5) and conclude (section 6). An appendix contains examples of the results obtained in English (A), French (B) and Russian (C).

2 Material

The biomedical domain has grown multiple terminologies of varying coverage, some of which have an international dissemination. The terminology that is considered as having the best coverage of clinical data (Chute *et al.*, 1996) is the Systematised Nomenclature of Human and Veterinary Medicine, also called *SNOMED International* (Côté *et al.*, 1993). SNOMED includes both preferred and synonym terms for a large part of its concepts (see table 1: PF = preferred, SY = synonym). Its full version, in English, contains over 150,000 medical terms. A complete translation of SNOMED in French or Russian is not yet available. Nevertheless, specialized subsets of this terminology have been prepared for some medical specialties; the Microglossary for Pathology contains over 12,500 terms, and has been translated into French (Côté, 1996) and Russian (Emelin *et al.*, 1995).

The International Classification of Diseases, maintained by the World Health Organization, is in wide usage in many countries. It is specialized in diagnoses, and has been translated in many languages including French (CIM-10, 1993). The “analytical” part of the current revision (tenth revision, ICD-10), contains about 11,000 terms (table 2). The clinical modification of the ninth revision (ICD-9-CM) is still in use in the USA.

Concept code	Type	English term
D2-01110	PT	Acute sinusitis, NOS
D2-01110	SY	Acute infection of nasal sinus, NOS
D2-01110	SY	Acute inflammation of nasal sinus, NOS
D2-01140	PT	Acute suppuration of nasal sinus
D2-01140	SY	Acute suppurative inflammation of nasal sinus
D2-01140	SY	Acute suppurative sinusitis

Table 1: Preferred and synonym English terms for SNOMED concepts D2-01110 and D2-01140.

Concept code	French term
J01	Sinusite aiguë
J010	Sinusite maxillaire aiguë
J011	Sinusite frontale aiguë
J012	Sinusite ethmoïdale aiguë
J013	Sinusite sphénoïdale aiguë
J014	Pansinusite aiguë
J018	Autres sinusites aiguës
J019	Sinusite aiguë, sans précision

Table 2: French terms for ICD-10 concept J01 and its descendants.

Our method relies on a thesaurus that has synonym terms. It also uses a reference list of words of the addressed domain. This list can be drawn from the same thesaurus or from some other source. The general shape of our experiments takes SNOMED as base thesaurus and draws the reference list from the SNOMED terms (or μglossary) augmented with the ICD terms (revision 9 or 10; no Russian ICD was available to us). All punctuation characters were considered as delimiters, and word forms containing digits were filtered out. All forms were converted to lower case. Table 3 gives the precise nature and figures for each of the three languages.

	Thesaurus			Reference list	
	Source	Terms	Synonym series	Source	Word forms
French	Snomed μglossary	12,555	2,344	Snomed μglossary + ICD-10 word forms	8,874
English	Full Snomed	128,855	26,295	Snomed + ICD-9-CM word forms	49,627
Russian	Snomed μglossary	13,462	2,636	Snomed μglossary	9,871

Table 3: Material for the three languages.

3 Learning morphological families from a thesaurus

Our base method proceeds in two steps. In the first step (section 3.1), it aligns pairs of morphologically related word forms in related terms of the thesaurus. In the second step (section 3.2), it learns morphological rules from these examples and applies them to the reference word list.

3.1 Identifying initial pairs of word forms

Thesauri provide relations between terms. In this paper we exploit the *synonymy* relation. We have also just started to examine other relations, namely hierarchical relations (*is-a* or *part-of*) and cross-reference relations (Zweigenbaum & Grabar, 2000b).

3.1.1 Aligning morphologically related word forms Given two synonym terms $\{T_1, T_2\}$, where each term is a set of word forms, we consider the pairs of word forms $\{IS_1, IS_2\}$ found in the two terms (with $IS_1 \in T_1$ and $IS_2 \in T_2$) and that share a "long enough" initial substring I. We worked with minimal lengths $\lambda = \text{length}(I)$ of three and four characters; this threshold is indeed a parameter of the method. Unless otherwise mentioned, in this paper, this threshold is set to 4. For instance, given the terms in table 1, this algorithm aligns the following word pairs:

(4) {sinus*itis,* sinus} *(D2-01110, D2-01140)*
{suppuratio*n,* suppurati*ve*} *(D2-01140)*

Note that such an initial common substring algorithm, applied in an unconstrained context (*e.g.*, our reference list of word forms), would give very bad results. For instance, a four-character initial substring length would collect together all the words that start with *abso-*: *absolute*, *absonum*, *absorbable*, *absorbent*, *absorber*, *absorptiometry*, *absorption*, *absorptive*. This would be correct for the *absorb-*/*absorp-* family, but would mix with them *absolute* and *absonum*. Also, starting from a specialized thesaurus, this method may identify domain-specific suffixes (*e.g.*, *-itis*, *-oma* or *-idae*) that a standard stemmer might not have in store.

Depending on the value of λ, the algorithm can also make mistakes. For instance, starting from the terms in table 1, a threshold of 3 instead of 4 also aligns the following word pair:

(5) {inf*ection,* inf*lammation*} *(D2-01110)*

3.1.2 Collecting morphological families Each word pair specifies that two word forms are linked by a morphological relation: they belong to the same morphological family. We can then try to recover these morphological families. Two word forms are included in the same family in the following situations:

- They were identified as a word pair by the alignment process;
- Both were segmented by the alignment process on the same maximal initial substring I;
- By transitivity, two families that share a common word form are merged.

For instance, the word pairs {ischia*l*, ischia*dic*} (T-16330) and {ischi*al*, ischi*um*} (T-12351) result in the morphological family *ischiadic, ischial, ischium* which is correctly separated from *ischaemia, ischaemic, ischemia, ischemic*, although their initial four-character substring (*isch-*) is the same.

3.1.3 Results This method was applied to the SNOMED Microglossary for Pathology (French and Russian) and to the full SNOMED (English). The results are displayed on table 4. The full English SNOMED, with a size ten times larger than that of the French or Russian Microglossary, yields six times more unique word pairs.

The precision of word pairs and morphological families was evaluated through human review. All of them were examined for French and Russian, whereas for English word pairs and families, random

Knowledge elements	French	English	Russian
Terms	12,555	128,855	13,462
Series of synonyms	2,344	26,295	2,636
Word pairs	1,572	15,549	2,445
Unique word pairs	1,086	6,556	1,535
Precision	99%	94.3±2.2%	99.9%
Suffix strings	446	2,499	676
Morphological families	623	3,188	773
Precision	97.9%	95.8±1.6%	99.9%
Words per family	2.53	2.90	2.74

Table 4: Initial pairs of word forms and morphological families, (λ = 4).

samples were extracted and reviewed (one every 15 word pairs, one every 5 families). A pair of word forms was considered correct if these forms were related through inflection ({*prosthesis*, *prostheses*}), derivation ({*cochlea*, *cochlear*}), compounding ({*atrophy*, *atrophoderma*}), or a combination thereof ({*lymphoid*, *lymphocytoma*}: derivation *-oid vs* compounding *-cytoma*). However, two word forms sharing only a general prefix or radical (*e.g.*, {*hypokalemia*, *hypomagnesaemia*}, {*autoeczematisation*, *autosensitization*}) were not considered a good match.

Appendix A.1 shows the most frequent suffix strings found in unique English word pairs: occurrences of domain-specific suffix strings can be found, *e.g.*, *-osis*, *-idae*, *-emia* and *-ectomy* (the latter two being components of compounds rather than suffixes). The most productive initial substrings and the first morphological families are also included in that appendix.

3.1.4 Discussion The precision of the word pairs and of the morphological families is extremely high, and does not decrease much when using a three-character threshold. We believe this is due to the very constrained context in which initial common substring matching is performed: synonym terms.

Errors occur when two word forms occurring in two synonym terms accidentally share a long enough initial substring. This is the case in synonyms terms for concept C-A0700: *Oral contraceptive preparation, NOS* and *Birth control pill, NOS* lead to the erroneous alignment of {*control*, *contraceptive*}.

It would be possible to create manually exception lists of word pairs, which would prevent errors produced at a low threshold. However, in these experiments, we have opted for a method that can be applied to another language or domain with virtually no human intervention (currently, the only human intervention is an optional tuning of the threshold).

The word pairs identified include the three types of morphological variation (inflection, derivation and compounding). Additional linguistic knowledge (lemmatization) may cancel the effect of inflection, so that only derivation and compounding remain. This will be shown later in this paper (section 4). The observation of occurrences of suffix strings in initial position in other word forms may help differentiate (compositional) radicals from derivational or inflectional suffixes; however, we have not yet come out with an automated method for separating out these three types of variation.

3.2 Inducing rules and applying them to reference word list

The word pairs aligned in the first step may be instances of general morphological relations that also apply to other word forms. To uncover these relations, these word pairs are considered as examples from which morphological rules are induced (see figure 1). These rules are potentially applicable to identify

other word pairs that undergo the same morphological relation.

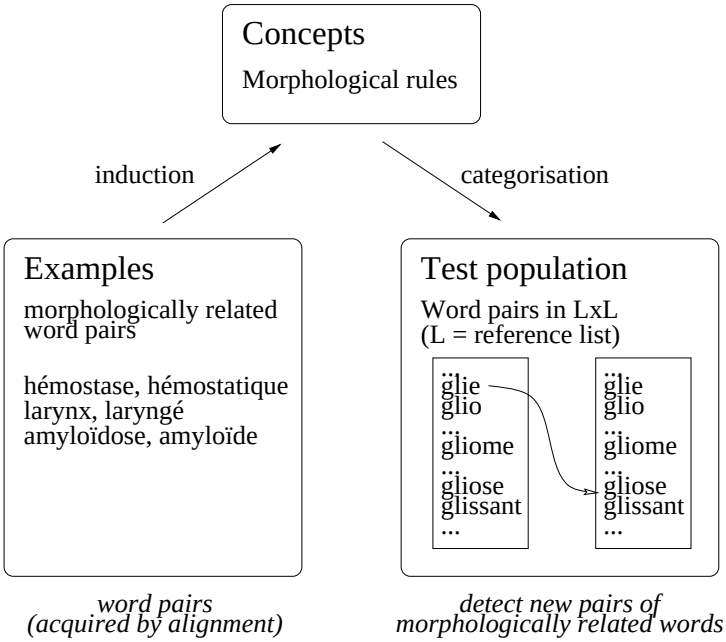

Figure 1: Inducing rules and applying them to the reference word list.

3.2.1 Inducing morphological rules We aim to obtain rules that embody the minimal specific features of the word pairs, while keeping a good generality. Each example $\{IS_1, IS_2\}$, where I is the maximal initial common substring of the two word forms, is generalized into a rule $\{S_1, S_2\}$ that can be interpreted as follows: given a word form that ends with suffix string S_1, one can derive a word form in which S_1 is substituted with S_2; or the reverse, these rules being considered symmetric. We note rules $\{S_1, S_2\}$ with the abbreviated notation $S_1|S_2$.

Concretely, aligning two word forms $\{IS_1, IS_2\}$ identifies at the same time the suffix strings $\{S_1, S_2\}$, thus the associated rule $S_1|S_2$. Some of the most frequent rules obtained for French are (ϵ denotes the null string):

$\epsilon|s$ (plural inflection) {*articulaire, articulaires*}
$\epsilon|e$ (feminine inflection) {*surrénal, surrénale*}
e|*ique* (adjectives ending with -*ique*) {*prostate, prostatique*}
e|*que* (adjectives ending with -*ique*) {*hyperkaliémie, hyperkaliémique*}
e|*aire* (adjectives ending with -*aire*) {*valvule, valvulaire*}
me|*sarcome* (derivation -*ome* / compounding -*osarcome*) {*mélanome, mélanosarcome*}

and for English:

$\epsilon|s$ (plural inflection) {*organ, organs*}
a|*osis* (-*osis* derivation) {*diverticula, diverticulosis*}
us|*osis* (-*osis* derivation) {*thrombus, thrombosis*}
a|*c* (-*ia* / -*ic* derivation) {*aphasia, aphasic*}
aemia|*emia* (alternation) {*tularaemia, tularemia*}
$\epsilon|al$ (-*al* derivation) {*orbit, orbital*}

Suffix strings are collected at the same time. For instance, suffix strings *-s*, *-a*, *-osis*, *-us*, etc. are identified in the word pairs cited in the above rules.

3.2.2 Identifying new word pairs We now use these rules to identify new word pairs in a larger reference list of word forms. This reference list was prepared as described in section 2.

For a given language, each morphological rule is applied to each word form W_i of the reference list, producing another word form W_j. If the resulting word form W_j also belongs to the reference list, *i.e.*, it is attested in the language, the word pair $\{W_i, W_j\}$ is considered as morphologically related (the implemented program actually uses a more efficient algorithm, based on a TRIE structure containing, in reverse form, all the words in the list). It is then added to a new set of word pairs. If the reference list contains all the word forms of the bootstrapping synonym terms, as is the case here, all the word pairs aligned at bootstrapping are found again here. The aim is that new word pairs can be identified too.

As in the alignment step, we limit noise by constraining the initial substring I of each pair to have a minimal length over a given threshold. We also set this threshold to four characters.

Here again, word pairs are merged into morphological families. These families are either extensions of the ones found in the initial step (when they contain a word pair that existed in the initial step) or are new families.

3.2.3 Evaluating precision and recall *Precision* was defined as the proportion of word pairs identified by our method that actually are in a relation of inflection, derivation, compounding or a combination thereof. Precision was evaluated by a human review of the word pairs and morphological families returned by our program. For English, the size of the results was too large, so that samples were extracted for review: one word pair every 15 and one morphological family every 5. We computed a confidence interval for these estimates with $\alpha = 0.05$.

For English, we are in a position where morphological resources for derivation (as well as inflection) are available: those distributed in the UMLS Specialist lexicon and its `lvg` tool (National Library of Medicine, 1999; McCray *et al.*, 1994). `lvg` can serve as an inflection and derivation generator, and has knowledge specific to the medical vocabulary. We ran `lvg` on the list of word forms collected for English to identify the word pairs where either (i) one word form was an inflected form of the other, or (ii) one word form was a derived form of the other (we ran it with options (i) `lvg -m -fi` to produce inflections and (ii) `lvg -m -fRf` for derivations). We took these word pairs as the gold standard for evaluating the *recall* of our method. Recall was defined as the proportion of `lvg`-returned word pairs also identified by our method.

3.2.4 Results The application of induced rules was performed on the material described above: for French, word forms from SNOMED Microglossary for Pathology and ICD-10; for English, word forms from full SNOMED and ICD-9-CM; for Russian, word forms from SNOMED Microglossary for Pathology. The results are shown on table 5.

The number of unique word pairs identified, compared with that obtained initially, is multiplied by 4. Morphological families are twice as numerous, and they are larger on average. Recall values for inflection and derivation are shown on table 6.

Example results are shown in appendices for English (A.2), French (B) and Russian (C).

3.2.5 Discussion Precision was still excellent after this induction step for French and Russian on both word pairs and families, and was good for English. The acquired rules only have an associative value:

Knowledge elements	French	English	Russian
From initial step			
Unique word pairs	1,086	6,556	1,535
Precision	99%	94.3±2.2%	99.9%
Morphological families	623	3,188	773
Words per family	2.53	2.90	2.74
Input			
Word forms	8,874	49,627	9,871
Output			
Morphological rules	567	3,039	834
Word pairs	4,573	22,372	6,399
Precision	98.3%	92.5±1.3%	99.6%
Morphological families	1,678	6,550	1,709
Precision	97.3%	91.9±1.5%	99.6%
Words per family	3.08	3.48	3.38

Table 5: Pairs of word forms and morphological families after induction (λ = 4).

Knowledge elements	Automatic	`lvg`	Recall
Word pairs	25,740		
Inflection		2,697	91.2%
Derivation		2,973	79.2%

Table 6: Inflection and derivation recall for English.

given two word forms, the rules can propose to relate them or not. It would not be wise however to apply these rules to any word form which ends with one of its suffix strings in order to generate new word forms. In particular, the rules that include the zero suffix ϵ do not mean that a correct word form can be obtained by applying them to any word form! By using these rules in an associative manner, *i.e.*, only on attested word forms, we limit the risk of producing impossible word forms or irrelevant stemming. The fact that the reference word list mostly contains domain-specific words may also have helped. Examples of errors include word pairs {*chin*, *china*} and {*plan*, *plane*}.

The results obtained with French and Russian are comparable, but cannot be matched directly with those for English because the size of the source material was much larger for English. In a complementary experiment (with λ = 3), we downsized English data by keeping one every 10 series of synonyms and saw that in this setting, the number of rules was less important for English (445) than for French (566).

Recall for English inflected word pairs was very good (over 90%; see table 6), and recall for derivation was fairly good too (close to 80%). Silence is basically due to rules that were not represented in the training material (word forms in SNOMED terms), or to rules that could have been observed in the training material but were not identified in synonym terms by the alignment method.

Our method identifies a much larger number of word pairs than were obtained by `lvg`. Several reasons can be invoked to account for this difference. First indeed, the evaluation of precision shows that 7.5±1.3% of these word pairs are spurious. Second, `lvg` did not generate the compound word forms found by our method. But, more important, our rules are somewhat redundant: one rule may perform the same substitution as the combination of several more elementary rules also identified on aligned word pairs. Given n words in a family, it can then collect up to $O(n^2)$ word pairs ($\frac{n(n-1)}{2}$), depending on the redundancy of the acquired rules.

In the most frequent rules, inflection again is the most frequent (*-s, -e, -es*); adjectival (*-ique, -aire*) and nominal (*-ation*) derivation comes next; domain-specific suffixes (*-ose*, *-ome*) and compound components (*-sarcome*) come later.

4 Using more initial linguistic knowledge

The above method uses no *a priori* linguistic knowledge. It is therefore potentially applicable to a large set of languages. It would be interesting to examine whether "low-level" linguistic knowledge, if available, might help this acquisition method.

On the one hand, morphological relations are generally constrained by the part-of-speech and morpho-syntactic features of the word forms involved. Therefore, one should expect higher precision if the same algorithms are applied to morpho-syntactically annotated input: that obtained with a *tagger* that assigns each word its part-of-speech (POS) and the value of some features such as number, gender, verb form, etc.

On the other hand, as mentioned in the introduction, inflectional morphology is well handled in many languages. Rather than collecting knowledge on inflectional, derivational and compounding variation altogether, we may therefore try to build on existing resources and, on the contrary, manage to cancel inflectional variation. Applying a *lemmatizer* to the input replaces each inflected form with its canonical form (lemma). Knowledge acquisition should then be able to focus on derivational and compounding relations, and maybe uncover more of these.

4.1 Starting from part-of-speech-tagged input

In this section, we introduce syntactic constraints on rules by applying the previous method to POS-tagged input: each word form in the thesaurus terms was tagged with its part-of-speech; and each word form in the reference word list was tagged with all its possible parts-of-speech. This experiment [POS] was performed on French, using a specifically trained version of the Brill tagger (Brill, 1995).

For the French [POS] experiment, we first used the Brill tagger trained for French at INaLF (Lecomte, 1998) to tag a subset of the Microglossary terms with the INaLF tagset. We then manually corrected the errors in the assigned tags, and trained the tagger on this corpus. The process was repeated on subsets of the thesaurus of increasing sizes, until we obtained a satisfactory POS-tagged version of the thesaurus. As the result of a tuning process (Lecomte, 1998), the tagset for French includes number features for nouns and adjectives, but no gender; the different verb forms considered are infinitive, past participle, present participle, and finite form.

Then all these tagged SNOMED word forms were copied into the reference word list, replacing the non-tagged SNOMED word forms therein. Word forms with several possible part-of-speech tags have several entries in this list. For instance, the French word *technique* may be either a noun (English *technique*) or an adjective (English *technical*). The remaining non-tagged word forms (ICD-10 forms not occurring in the Microglossary) were tagged manually.

Following a common convention, part-of-speech tags were appended to words or suffixes with a slash character "/" as separator. For instance, one term for code M-09150 is *tissu/SBC:sg égaré/ADJ:sg pendant/PREP la/DTN:sg manipulation/SBC:sg technique/ADJ:sg*. We handle these tags as extensions to word suffix strings, so that the rule *e/SBC:sg|ique/ADJ:sg* stands for a rule *e|ique* constrained to apply between a singular noun (SBC:sg) and a singular adjective (ADJ:sg).

4.2 Starting from lemmatized input

A lemmatized form of the terms and word list was prepared using Namer's FLEMM lemmatizer (Toussaint *et al.*, 1998). FLEMM takes as input words tagged with the Brill-INaLF tagset, and computes for each of them the corresponding lemma; it also refines the tags by adding, *e.g.*, values for gender feature where possible. This experiment [LEM] was performed on French. FLEMM was run on the tagged terms and on the tagged word list, and tagged word forms were replaced with their root forms.

Finally, input containing lemmata with part-of-speech tags was also prepared for French [POS-LEM]. Tags with morphosyntactic features (number, gender, verb forms) were updated to tags more relevant to canonical forms, where these morphosyntactic features were erased.

4.3 Results

The learning method was applied to each of these inputs.

	[STD]	[POS]	[LEM]	[POS-LEM]
Terms	12,555	12,555	12,555	12,555
Series of synonyms	2,344	2,344	2,344	2,344
Word pairs	1,572	1,597	1,414	1,433
Unique word pairs	1,086	1,104	935	948
Precision	99 %	99 %	98.3 %	98.3 %
Morphological rules	567	594	507	522
Reference word forms	8,874	8,989	7,187	7,277
Word pairs	4,573	4,411	2,390	2,389
Precision	98.3 %	98.9 %	97.3 %	97.9 %
Morphological families	1,678	1,647	1,064	1,078
Precision	97.3 %	98.3 %	96.4 %	97.1 %
Words per family	3.08	3.06	2.79	2.77

Table 7: Influence of POS-tagging and lemmatization (French, $\lambda = 4$).

Table 7 shows the results obtained for the three experiments; results with no *a priori* linguistic knowledge [STD] are repeated for comparison. Word counts (word pairs, word forms, words per family) are given for tagged word forms in the [POS] experiment, for lemmatized words in the [LEM] experiment and for tagged, lemmatized words in [POS-LEM].

In the [POS] experiment, the number of aligned word pairs has increased compared to the previous, non-tagged experiment. This comes from polycategorial words, which occur with different part-of-speech tags in different contexts. On the one hand, they can undergo part-of-speech *conversions* (such as the word *technique* mentioned earlier). On the other hand, the same pair of untagged word forms may give rise to several pairs of tagged word forms when one of them is polycategorial. The number of induced rules is increased accordingly. The final volume of word pairs and families is nevertheless decreased, as a result of syntactic constraints on rule application. Incorrect word pairs are more eliminated than correct ones, as is reflected by the increase in precision. Note however that the variations are very small (less than 1%).

In the [LEM] experiment, all these figures are lower than those obtained in [STD]. As can be expected, the volume of initial word pairs and rules is reduced (-14% of unique initial word pairs, 10% of the rules) since inflected variants have been suppressed. The rules removed are those that only involved inflection (*e.g.*, feminine gender inflection $e|\epsilon$), or were variant versions of derivation or compounding rules in-

cluding inflectional variations (*e.g.*, feminine gender inflection combined with *-al* adjectival derivation *e|al*, yielding variant rule *e|ale*). This modest reduction in rules produces a much larger reduction in the number of final word pairs (–48%) and families (–37%). The reduction in word pairs may be explained by the frequency of application of rules combining inflection and other morphological operations; the reduction in morphological families comes from the fact that many families only contained inflectional variants. The precision of [LEM] is lower than that of [STD], but here again the variation is very small (less than 1%).

The results of the [POS-LEM] experiment are consistent with what was observed for [POS] and [LEM] independently. Compared with [LEM] alone, initial word pairs are slightly more numerous because of multiple categorizations. The number of final word pairs and families is however nearly unchanged: syntactic filtering compensates for the preceding increase. It also reduces the number of errors, so that precision is increased and nearly reaches that of [STD]. Compared with [POS] alone, the same pattern as from [STD] to [LEM] can be observed.

Example results for [POS-LEM] can be found in appendix B.2.

4.4 Discussion

In the [POS] experiment, a pair of untagged word forms identified in the [STD] experiment may be filtered out for different reasons.

1. Remaining tagging errors: a tagged word form was not identified because it was incorrectly tagged;
2. Missing category in reference word list: a polycategorial word form lacked some entries in the tagged word list, so that a rule that would have applied to its non-tag part could not apply because of tag discrepancy; this is the case, *e.g.*, of the pair {*gris, grise*} (*gray* masculine, feminine): *gris* is ambiguous between singular and plural, but only attested as plural *gris/ADJ:pl* in the reference list, so that the feminine gender inflection rule *ϵ/ADJ:sg|e/ADJ:sg* cannot apply to it.
3. Missing tagged rule variant: here, a rule was learnt with a specific syntactic constraint represented in the aligned word forms; however, these constraints are more specific than would be appropriate, which blocks the application of this rule in a correct context; for instance, the plural rule *l/ADJ:sg|ux/ADJ:pl* could not apply to the pair of noun forms *canal/SBC:sg* and *canaux/SBC:pl*, although a similar rule (not learnt from the aligned word pairs) exists in French for nouns: *l/SBC:sg |ux/SBC:pl*;
4. Rules learnt on very specific examples, which by chance happened to apply to other specific examples. For instance, rule *ϵ|ose* was learnt on {*hyalin, hyalinose*}. When tagged, the rule becomes *ϵ/ADJ:sg|ose/SBC:sg*, and does not apply anymore to {*cholestérol/SBC:sg, cholestérolose/SBC:sg*}.
5. Correct blocking of a morphologically unrelated word pair; for instance, rule *ϵ/ADJ:sg|e/ADJ:sg* produces the feminine form of regular adjectives; the adjective constraint blocked the incorrect pairs {*gain/SBC:sg, gaine/SBC:sg*} ({*gain, sheath*}) as well as {*médecin/SBC:sg, médecine/ADJ:sg*} ({*physician, medicine*}); the same process went on for plural rule *ϵ/SBC:sg|x/SBC:pl*, which was blocked on {*verte/ADJ:sg, vertex/SBC:sg*} ({*green, vertex*}) and {*simple/ADJ:sg, simplex/SBC:sg*}.

In summary, using part-of-speech information decreases the error rate (increases precision), but also decreases recall. Some of these problems might be reduced by considering that a polycategorial word form occurrence should represent all its categories, as proposed by Krovetz (1993) in a different context.

Working on lemmatized input sensibly reduces the number of word pairs and morphological families by filtering out inflection-related rules. The resulting rules, as expected, are focused on derivation and compounding, which are the semantically more interesting phenomena, as can be seen on the most frequently applied rules and the most frequent suffix strings (appendix B.2, [POS-LEM], French). The number of errors is kept low; but as the number of word pairs is much lower, the precision is mechanically decreased.

5 Synthesis

The rationale behind the methods in this work is that terminologies contain a wealth of linguistic knowledge that can be "reverse engineered" to a large extent. Synonymy relations include explicit morphological variants for *terms*, purposely inserted to facilitate "natural language" indexing and search. They also contain more implicit morphological variants for *single words*. Observing such morphological word variants in the synonym terms that denote the same concept is a very strong clue that these words belong to the same morphological family. We have seen, with experiments in three different languages (of the romance, germanic and slavic families), that this can lead to high precision collection of morphologically related word pairs.

This initial collection is then used to bootstrap a larger acquisition process that identifies more morphologically related word pairs across concepts. This increases recall, while keeping a very good precision. The application to large-size English data showed that recall can be fairly good too. Its more precise dependence to training set size needs to be investigated.

If part-of-speech-tagged input is available, rules can be made syntactically more precise and block some incorrect word pairs; a slight decrease in recall is incurred because of some polycategorial word forms. If, furthermore, lemmatized input can be obtained, rules and affixes are "cleaned" from the influence of inflection. The interesting affixes, among which morphemes can be found, are those of derivation and compounding, which are then more overtly exhibited.

No distinction was made in this work between derivational affixes and compound components. Such a distinction seems important for information retrieval, as keeping only the first component of a compound word looses information. More work is needed to learn how to make it automatically.

In any case, a (small) proportion of the word pairs identified still require human correction. They could be handled as exceptions, as is the case in many morphological processors such as `lvg` and FLEMM.

6 Conclusion

We proposed a method for learning morphological data for a given language and domain, through the use of synonym series of a thesaurus. This very specific semantic context yields a very good precision (97.3% for French, 91.9±1.5% for English and 99.6% for Russian morphological families). Recall could be evaluated for English; this showed that with the large training sample available for English (26,295 synonym series), 91.2% of inflection variations and 79.2% of the derivation variations occurring in the 49,627-word list were covered. We also showed that additional constraints using syntactic knowledge can improve precision further, with a limited loss in recall.

By relying on a thesaurus, this method can complement the panel of approaches cited above and help obtain domain-specific morphological data, which can be used to complete or replace traditional stemming methods.

Experiments have been started to examine the productivity of other thesaural relations such as hyper-

onymy (Zweigenbaum & Grabar, 2000b). This will allow us to work with terminologies that do not have synonym terms but are hierarchically organized, such as the International Classification of Diseases.

Acknowledgments

We wish to thank Dr Roger Côté for a precommercial version of the French SNOMED Microglossary for Pathology, Yvan Emelin for a draft version of the Russian SNOMED Microglossary for Pathology, INaLF for lending us a version of Brill's tagger trained for French, Fiametta Namer for the FLEMM lemmatizer, and the NLM for the `lvg` tool, Specialist Lexicon and UMLS Metathesaurus from which data for the English versions of SNOMED and ICD were obtained.

References

Brill, E. (1995). Transformation-based error-driven learning and natural language processing: A case study in part-of-speech tagging. *Computational Linguistics*, **21**(4), 543–565.

Burnage, G. (1990). *CELEX - A Guide for Users*. Nijmegen: Centre for Lexical Information, University of Nijmegen.

Chute, C. G., Cohn, S. P., Campbell, K. E., Oliver, D. E. & Campbell, J. R. (1996). The content coverage of clinical classifications. *Journal of the American Medical Informatics Association*, **3**(3), 224–233. for the Computer-Based Patient Record Institute's Work Group on Codes and Structures.

CIM-10 (1993). *Classification statistique internationale des maladies et des problèmes de santé connexes — Dixième révision*. Organisation mondiale de la Santé, Genève.

R. A. Côté, D. J. Rothwell, J. L. Palotay, R. S. Beckett & L. Brochu, Eds. (1993). *The Systematised Nomenclature of Human and Veterinary Medicine: SNOMED International*. Northfield: College of American Pathologists.

Côté, R. A. (1996). *Répertoire d'anatomopathologie de la SNOMED internationale, v3.4*. Université de Sherbrooke, Sherbrooke, Québec.

Dal, G., Namer, F. & Hathout, N. (1999). Construire un lexique dérivationnel : théorie et réalisations. In P. Amsili, Ed., *Actes de TALN 1999*, Cargèse.

Emelin, I. V., Levenson, R., Perov, Y. L. & Rykov, V. V. (1995). A Russian version of SNOMED-International. In R. A. Greenes, H. E. Peterson & D. J. Protti, Eds., *Proceedings of the 8 th World Congress on Medical Informatics*, pp. 173–173, Vancouver.

Grabar, N. & Zweigenbaum, P. (1999). Acquisition automatique de connaissances morphologiques sur le vocabulaire médical. In P. Amsili, Ed., *Actes de TALN 1999*, pp. 175–184, Cargèse.

Jacquemin, C. (1997). Guessing morphology from terms and corpora. In *Actes, 20th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'97)*, pp. 156–167, Philadelphia, PA.

Krovetz, R. (1993). Viewing morphology as an inference process. In *Proceedings of the 16th Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval*, pp. 191–202.

Lecomte, J. (1998). *Le catégoriseur Brill14-JL5 / WinBrill-0.3*. Technical report, INaLF. Available at `http://jupiter.inalf.cnrs.fr/WinBrill/winbrill.etiquetage.html`.

McCray, A. T., Srinivasan, S. & Browne, A. C. (1994). Lexical methods for managing variation in biomedical terminologies. In *Proceedings of the 18th Annual SCAMC*, pp. 235–239, Washington: Mc Graw Hill.

National Library of Medicine (1999). *UMLS Knowledge Sources Manual*. National Library of Medicine.

Schulz, S., Romacker, M., Franz, P., Zaiss, A., Klar, R. & Hahn, U. (1999). Towards a multilingual morpheme thesaurus for medical free-text retrieval. In *Proceedings of MIE'99*, Ljubliana, Slovenia: IOS Press.

Silberztein, M. (1993). *Dictionnaires électroniques et analyse automatique de textes : le système INTEX*. Paris: Masson.

Toussaint, Y., Namer, F., Daille, B., Jacquemin, C., Royauté, J. & Hathout, N. (1998). Une approche linguistique et statistique pour l'analyse de l'information en corpus. In P. Zweigenbaum, Ed., *Actes de TALN 1998*, Paris.

Wingert, F., Rothwell, D. & Côté, R. A. (1989). Automated indexing into SNOMED and ICD. In J. R. Scherrer, R. A. Côté & S. H. Mandil, Eds., *Computerised Natural Medical Language Processing for Knowledge Engineering*, pp. 201–239. Amsterdam: North-Holland.

Wolff, S. (1987). Automatic coding of medical vocabulary. In N. Sager, C. Friedman & M. S. Lyman, Eds., *Medical Information Processing - Computer Management of Narrative Data*, chapter 7, pp. 145–162. Reading Mass: Addison Wesley.

Xu, J. & Croft, B. W. (1998). Corpus-based stemming using co-occurrence of word variants. *ACM Transactions on Information Systems*, **16**(1), 61–81.

Zweigenbaum, P. & Grabar, N. (2000a). Expériences d'acquisition automatique de connaissances morphologiques par amorçage à partir d'un thésaurus. In *Proceedings of the 12 th Conference RFIA*, pp. II–101–II–110, Paris, France: AFCET.

Zweigenbaum, P. & Grabar, N. (2000b). Liens morphologiques et structuration de terminologie. Submitted to *IC 2000 : Ingénierie des connaissances*.

A Example results for English

A.1 Initial results

Most frequent suffix strings found in unique English word pairs (frequency over 100):

(6)

652 -s	*370 -al*	*206 -y*	*130 -ed*	*109 -emia*
624 -a	*359 -e*	*193 -idae*	*130 -c*	*108 -tic*
543 -osis	*269 -ic*	*163 -es*	*125 -ectomy*	*106 -ia*
512 -us	*238 -um*	*153 -is*	*111 -sis*	*105 -ing*

Most productive initial substrings (frequency over 10):

(7)

60 hyper-	*17 post-*	*14 pleur-*	*13 bronch-*	*11 thyro-*
29 ureter-	*17 fibro-*	*14 phospho-*	*13 arter-*	*11 radio-*
28 hypo-	*16 phosph-*	*14 neur-*	*12 sulf-*	*11 gastr-*
27 esophag-	*15 trans-*	*14 coccy-*	*12 meth-*	*11 derm-*
25 oesophag-	*15 neuro-*	*14 cardi-*	*12 lymph-*	*11 acet-*
22 anti-	*15 cyst-*	*14 aort-*	*12 colo-*	
21 osteo-	*14 urethr-*	*13 vagin-*	*12 arteri-*	

First 20 morphological families (in alphabetical order):

(8) *abdomen, abdominal, abdominalis, abdominoplasty, abdominocentesis;*
abducens, abducent;
abetalipoproteinemia, abetalipoproteinaemia;
ablepharia, ablepharon;
abnormal, abnormality, abnormalities;
abomasal, abomasum;
abortive, abortivum;
abortient, aborticide, abortifacient;
aborter, abortus, abortion, abortions;
abrasion, abrasions;
abscess, abscesses;
abscisic, abscisin, abscissin;
absent, absence;
acanthamoeba, acanthamoebidae, acanthamoebosis, acanthamoebiasis, acanthamoebiidae;
acanthocephalus, acanthocephalosis;
acanthocheilonema, acanthocheilonemiasis;
acanthosis, acantholysis;
acanthurus, acanthuridae;
acanthocyte, acanthrocyte;
acarapis, acarapisosis.

A.2 Results after induction

The most frequent rule applications are:

(9)

1285 ϵ|s	*218 ϵ|ing*	*163 ϵ|us*	*118*	*109 ϵ|ic*
362 a|us	*214 ϵ|is*	*155 ed|ing*	*ectomy|otomy*	*105 e|is*
330 ϵ|e	*193 ϵ|ed*	*147 ϵ|es*	*118 a|idae*	*103 on|ve*
290 ϵ|i	*192 ϵ|al*	*144 a|c*	*115 ϵ|d*	*103 i|us*
281 a|um	*188 a|osis*	*123 ϵ|er*	*115 ate|ic*	*101 osis|us*
243 m|s	*166 ϵ|l*	*123 a|e*	*113 ϵ|y*	
225 ϵ|a	*166 idae|us*	*120 ed|ion*	*112 e|us*	

2498 suffixes are involved, among which the most frequent are:

(10)

2484 -s	*781 -ing*	*401 -o*	*319 -m*	*220 -ation*
2247 -a	*761 -ed*	*386 -c*	*308 -ase*	*213 -oplasty*
1970 -us	*642 -ectomy*	*367 -l*	*262 -n*	*211 -ine*
1696 -e	*613 -y*	*365 -otomy*	*252 -ia*	*211 -d*
1233 -al	*600 -i*	*364 -ate*	*245 -er*	*210 -tic*
989 -um	*584 -is*	*361 -ion*	*227 -yl*	
984 -ic	*531 -es*	*348 -itis*	*225 -on*	
921 -osis	*484 -idae*	*335 -sis*	*221 -ar*	

10,942 initial strings are involved; the most frequent ones are (frequency over 20):

(11)

68 hyper-	*36 limb-*	*29 radi-*	*25 dent-*	*22 phospho-*
58 ureter-	*36 esophag-*	*29 jejun-*	*25 brom-*	*22 duoden-*
55 sept-	*35 rect-*	*29 carp-*	*25 aort-*	*22 derm-*
50 enter-	*33 chlor-*	*29 cardi-*	*24 scler-*	*21 tympan-*
49 cyst-	*32 vagin-*	*28 nephro-*	*24 choledoch-*	*21 trans-*
48 osteo-	*32 urethr-*	*28 gastr-*	*24 anti-*	*21 spin-*
43 fibro-	*32 arteri-*	*27 rhin-*	*24 adeno-*	*21 hepatic-*
40 neuro-	*31 hypo-*	*27 myelo-*	*23 phosph-*	*21 crani-*
38 neur-	*31 form-*	*26 pleur-*	*22 ventricul-*	*21 colo-*
36 oesophag-	*30 hepat-*	*25 medi-*	*22 plan-*	*21 acet-*

The first 20 morphological families are:

(12) *abdomen, abdomino, abdominal, abdominis, abdominalis, abdominopelvic, abdominoplasty, abdominocentesis;*
abducens, abducent;
abductor, abduction;
aberrant, aberration;
abetalipoproteinemia, abetalipoproteinaemia;
ablatio, ablation;
ablepharia, ablepharon;
abnormal, abnormis, abnormalis, abnormally, abnormality, abnormalities;
abomasal, abomasum, abomasitis, abomasopexy, abomasectomy;
abortient, aborticide, abortifacient;
aborter, abortus, abortion, abortive, abortions, abortivum;
abras, abrasion, abrasions;
abrupt, abruptio;

abscess, abscesses, abscessus, abscessation;
abscisic, abscisin, abscissin;
absent, absence;
absorption, absorptive;
abstract, abstracting;
abyssinian, abyssinica;
academy, academic.

B Example results for French (after induction)

B.1 Without linguistic knowledge [STD]

Most frequent rule applications:

(13)

1140 ϵ|s	*64 e|que*	*42 e|o*	*36 e|ome*	*33 ϵ|le*
290 ϵ|e	*60 e|s*	*42 al|o*	*36 aire|es*	*32 e|é*
143 ϵ|es	*55 ation|é*	*40 le|ux*	*34 ϵ|ment*	*31 me|sarcome*
74 e|ique	*53 se|x*	*40 e|ose*	*34 f|ve*	*31 atose|e*
67 aire|e	*43 l|ux*	*38 ite|o*	*33 ϵ|ne*	

Most frequent suffixes (total 447):

(14)

1254 -s	*143 -ome*	*80 -ation*	*50 -oïde*	*34 -tique*
1046 -e	*141 -ose*	*79 -sarcome*	*41 -euse*	*34 -ion*
268 -es	*114 -ux*	*63 -eux*	*39 -ale*	*33 -cytaire*
243 -o	*101 -al*	*57 -ve*	*38 -ée*	*32 -tion*
221 -é	*99 -que*	*56 -on*	*38 -ant*	*32 -ienne*
212 -aire	*97 -x*	*54 -ite*	*35 -um*	*31 -ement*
169 -me	*83 -ïde*	*51 -ne*	*35 -ment*	*31 -blastique*
166 -se	*81 -f*	*51 -l*	*35 -atose*	
165 -ique	*80 -le*	*51 -blastome*	*34 -us*	

Most frequent initial strings (total 2879):

(15)

29 myélo-	*21 adéno-*	*15 angio-*	*12 méning-*	*11 hyalin-*
25 ostéo-	*19 lympho-*	*13 plasmocyt-*	*12 mélano-*	*11 histiocyt-*
22 fibro-	*16 lipo-*	*13 immun-*	*12 lymphocyt-*	
22 fibr-	*16 granul-*	*13 chondro-*	*11 hémangio-*	

B.2 With linguistic knowledge [POS-LEM]

Most frequently applied rules:

(16)

45 aire/ADJ|e/SBC	*25 ion/SBC|oire/ADJ*	*17 e/SBC|é/ADJ*
39 e/SBC|que/ADJ	*25 al/ADJ|o/PFX*	*17 e/SBC|o/PFX*
38 me/SBC|sarcome/SBC	*24 ation/SBC|é/ADJ*	*17 ϵ/ADJ|ϵ/SBC*
34 e/SBC|ique/ADJ	*22 gne/ADJ|n/ADJ*	*16 atose/SBC|e/SBC*
32 ateux/ADJ|e/SBC	*20 ϵ/SBC|ux/ADJ*	*15 se/SBC|tique/ADJ*

14 sarcome/SBC|ïde/ADJ	*12 eux/ADJ|ose/SBC*	*11 blastome/SBC|me/SBC*
13 e/SBC|oïde/ADJ	*11 ique/ADJ|o/PFX*	*11 aire/ADJ|o/PFX*
13 al/ADJ|um/SBC	*11 f/ADJ|on/SBC*	*10 ement/SBC|é/ADJ*
13 al/ADJ|e/SBC	*11 carcinome/SBC|me/SBC*	*10 blastique/ADJ|ïde/ADJ*

Most frequent suffix strings (total 427):

(17)

209 -e/SBC	*45 -є/ADJ*	*27 -ation/SBC*	*15 -cytaire/ADJ*
88 -é/ADJ	*44 -se/SBC*	*22 -ien/ADJ*	*15 -carcinome/SBC*
82 -o/PFX	*44 -ome/SBC*	*20 -us/SBC*	*14 -s/SBC*
71 -ique/ADJ	*40 -ose/SBC*	*20 -um/SBC*	*13 -x/SBC*
68 -aire/ADJ	*36 -que/ADJ*	*18 -є/PFX*	*13 -ie/SBC*
61 -al/ADJ	*35 -eux/ADJ*	*17 -tique/ADJ*	*13 -blastique/ADJ*
59 -є/SBC	*30 -ïde/ADJ*	*17 -e/ADJ*	
58 -me/SBC	*28 -sarcome/SBC*	*16 -blastome/SBC*	

C Example results for Russian (after induction)

The first 40 rule applications for Russian (example (18)) only involve inflections except the three rules *ие|ный*, *є|ная* and *є|ный*, which all deal with adjectival derivation.

(18)

460 ая|ый	*103 ая|ого*	*74 го|е*	*52 ая|ых*	*38 ого|ые*
243 е|й	*98 ая|о*	*73 го|й*	*51 є|ов*	*37 є|м*
212 є|а	*97 о|ый*	*69 є|е*	*51 є|ная*	*36 й|м*
207 ая|ое	*96 ого|ый*	*67 е|я*	*50 ой|ые*	*35 ий|ое*
196 ая|ой	*96 а|ы*	*62 є|й*	*50 є|ы*	*34 є|ом*
186 ое|ый	*86 ая|ий*	*58 є|ный*	*49 й|я*	*32 о|ые*
140 ая|ые	*81 и|я*	*53 ое|ые*	*49 а|и*	*32 а|о*
113 ой|ый	*79 й|х*	*53 ие|ный*	*40 у|ь*	*32 а|ная*

Among the most frequent suffixes (over 100 occurrences; total 676; example (19)), *-ная* and *-ный* concern derivation and *-ма* is a compound component (English *-oma*).

(19)

1464 -ая	*486 -е*	*313 -ые*	*181 -ная*	*104 -ов*
1011 -ый	*398 -ой*	*279 -ы*	*163 -ие*	
625 -а	*342 -о*	*277 -и*	*162 -ий*	
569 -й	*335 -я*	*220 -ный*	*159 -ых*	
536 -ое	*314 -ого*	*182 -го*	*140 -ма*	

The most frequent initial strings (total 2,937, (20)) involve both loan word components (*фибро-* = *fibro-*) and Russian components (*почечн-*).

(20)

36 фибро-	*22 миело-*	*19 костн-*	*18 нейро-*
28 лимфо-	*21 фиброз-*	*19 кожн-*	*18 кишечн-*
24 почечн-	*21 легочн-*	*19 злокачественн-*	*17 остео-*

17 нерв-	*15 адено-*	*13 клапан-*	*11 хирургическ-*
17 диффузн-	*14 эндометри-*	*12 открыт-*	*11 фиброзн-*
17 врожденн-	*14 тяжел-*	*12 мембран-*	*11 тубулярн-*
17 базальн-	*14 полов-*	*12 липо-*	*11 сосудист-*
16 эпител-	*14 печеночн-*	*12 лимф-*	*11 слизист-*
16 хромосом-	*14 мышечн-*	*11 эндокринн-*	*11 серозн-*
15 клеточн-	*14 артери-*	*11 хроническ-*	*11 сердечн-*

Among the first 10 families (21), the eighth one contains the derived adjective *адвентициальная* whereas the *адено* family involves inflection, derivation and many compound forms.

(21) *абдоминальная, абдоминальный;*
аборт, аборта;
абсолютное, абсолютный;
абсцесс, абсцесса, абсцессом;
агрегат, агрегаты;
агент, агента;
агрессивная, агрессивный;
адвентиции, адвентиция, адвентициальная;
адгезии, адгезий, адгезия;
адено, аденоз, аденозы, аденома, аденомы, аденоматоз, аденолипома, аденофиброз, аденосаркома, аденосаркомы, аденофиброма, аденоакантома, аденоматозная, аденоматозном, аденоматозные, аденоматозный, аденоматозных, аденокарцинома, аденокистозный, аденоматоидная, аденосквамозный, аденокарциноидная, аденоамелобластома, аденофиброматозная;